How do economic and social factors influence national performance in the Olympic Games?

Statistical Methods for Official Statistics

Master's Degree in Data Science

Riccardo Corrente (1964746)

Academic Year 2024/2025

SAPIENZA
UNIVERSITÀ DI ROMA

Table of Contents

The independent variables

1 Goal of the analysis

- The primary objective of this analysis is to investigate how **economic and social factors** influence a country's performance in the Olympic Games.
- Specifically, the analysis focuses on five different indices (independent variables):
  - **GDP**: Countries with higher GDP generally have more resources to invest in sports, potentially leading to better performances.
  - **GDP per capita**: Indicates the average income per person.
  - **Population**: A larger population increases the potential talent pool, which can lead to more medal opportunities.
  - **HDI** (Human Development Index): A summary measure of average achievement in key dimensions of human development.
  - **Host nation**: A dummy variable indicating whether a country is hosting that specific edition of the Olympic Games or not.

The target variable

1 Goal of the analysis

The target variable in this analysis is:

- **Total number of Olympic medals:** This represents the overall success of a country in the Olympic Games, accounting for gold, silver, and bronze medals won across all Olympic events.

The total medal count serves as an indicator of athletic success and is hypothesized to be influenced by the independent variables, such as GDP per capita, total GDP, population, HDI, and host nation.

The final objective of this analysis is to evaluate how effectively these variables can **predict** a country's Olympic performance and identify the most influential factors.

Table of Contents

2 The construction process of the Dataset

The sources

2 The construction process of the Dataset

To build the final dataset, I had to merge different datasets from various sources. In particular:

- From the World Bank site, I gained the data about the *GDP* and the *Population*.
- The Olympic DataSet was imported from *Kaggle* and contains the list of all the athletes participating in the Summer Olympic Games from Athens 1896 to Tokyo 2020.
- The HDI Index was found in the *HDR* (Human Development Reports) website.

The Olympic Dataset

2 The construction process of the Dataset

- The Olympic Dataset has 250.832 rows, each one corresponds to an individual athlete competing in an individual Olympic event (athlete-events). The columns are the following:

	player_id	Name	Sex	Team	NOC	Year	Season	City	Sport	Event	Medal
252560	4986655	Sefora Ada	F	Equatorial Guinea	GEQ	2024	Summer	Paris	Athletics	Women's 100m	No medal
252561	9460001	Emanuela Liuzzi	F	Italy	ITA	2024	Summer	Paris	Wrestling	Women's Freestyle 50kg	No medal
252562	1972077	Isayah Boers	M	Netherlands	NED	2024	Summer	Paris	Athletics	4 x 400m Relay Mixed	Gold
252563	1899865	Kevin Staut	M	France	FRA	2024	Summer	Paris	Equestrian	Jumping Team	Bronze
252564	1924402	Charlie Carvell	M	Great Britain	GBR	2024	Summer	Paris	Athletics	Men's 4 x 400m Relay	Bronze

Cleaning process

2 The construction process of the Dataset

To create the final DataFrame with the total number of medals per country and year, the following steps were performed:

1. Filtered the dataset to include only medalist athletes.
2. Fixed data errors (e.g., ”Italy-1”) and ensured accurate team vs. individual medal counts.
3. Grouped data by Team and Year to calculate medals per country per Olympic Games.
4. Created a pivot table to track male and female participants, adding a column for the total number of athletes per country.

Merging datasets - Challenges

Merging the four datasets required addressing a key issue:

- The Olympics dataset uses the identifier **NOC** (National Olympic Committee), which differs from the **Code** used by the World Bank.
- To resolve this, I imported a mapping from Wikipedia to link **NOC** to **Country Name**.
- I standardized country names across all datasets, creating a unified key for merging.

Table of Contents

3 Hypothesis

Hypothesis

3 Hypothesis

- In order to compare the data from different years, I decided to analyze the data series by normalizing the variables per year.
- The final DataFrame includes additional columns such as `medals_norm`, `gdp_norm`, and `population_perc`. These variables naturally increase over the years, so the best way to evaluate their significance is to standardize them, otherwise the analysis could lead to biased results.

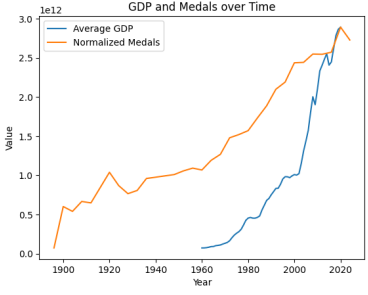


Table of Contents

4 Exploratory Data Analysis

Top countries by medals

4 Exploratory Data Analysis

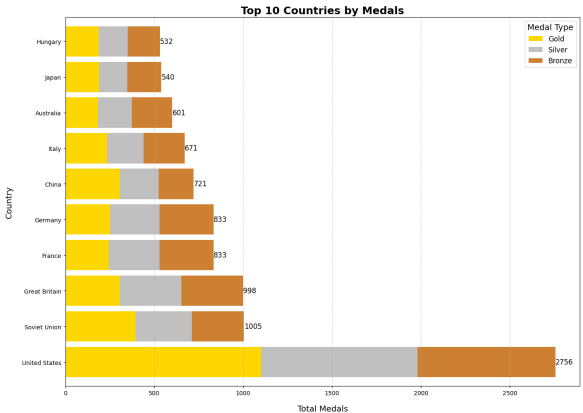


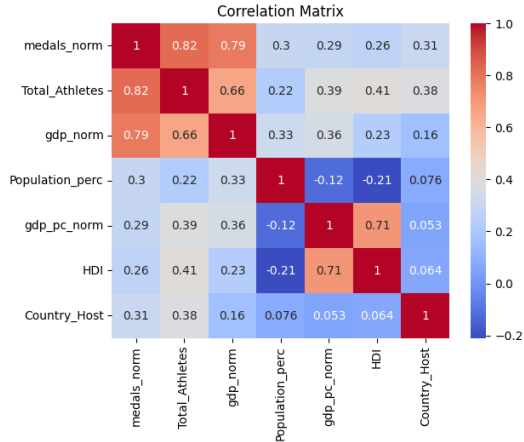
Figure: Top 10 countries by total medals in the history.

- The United States dominates the Olympic medal count with a significant lead, earning a total of 2,756 medals.
- The Soviet Union and Great Britain follow, with total medals slightly exceeding 1,000.
- Countries like Hungary and Australia, despite smaller populations, have consistently strong results, reflecting the importance of strong sports traditions despite smaller populations.

Correlation Matrix and Analysis

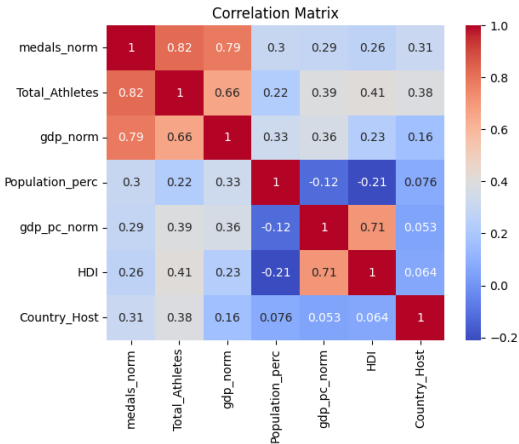
4 Exploratory Data Analysis

- **Total medals** shows a strong positive correlation with **Total Athletes** ($r = 0.82$) and **GDP** ($r = 0.69$), indicating that larger delegations and wealthier nations are associated with higher medal counts.
- Surprisingly a small correlation is observed between **GDP per capita** and **total medals** ($r = 0.22$), this is because of the presence of small countries with very high GDP per capita.
  - This is highlighted also by the negative correlation between the Population and the GDP per capita ($r = -0.12$).



Correlation Matrix and Analysis

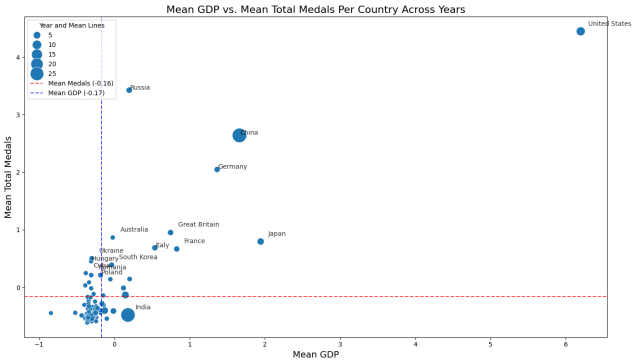
4 Exploratory Data Analysis



- Interestingly, **Population** shows only a weak correlation with **total medals** ($r = 0.35$), indicating that population size alone is not a key determinant of Olympic success.
- The correlation between **HDI** and **Total Medals** is small as well, reflecting that human development factors may not directly translate to Olympic success by themselves.

GDP & Population vs. Total Medal across years

4 Exploratory Data Analysis

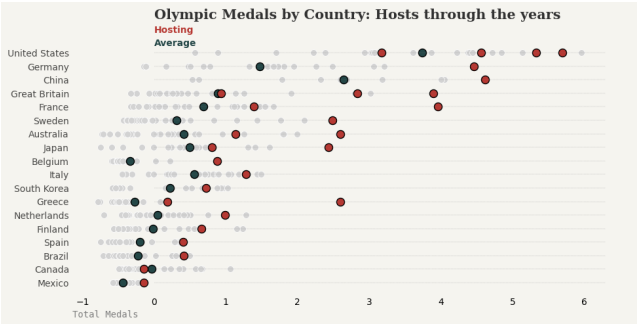


- Countries with high GDP (e.g., the United States and China) tend to win more medals.
- The size of the bubbles highlights population effects, with countries like India having larger populations but fewer medals.
- Mean lines (red for medals, blue for GDP) divide countries into four quadrants, helping to categorize their performance.

Does hosting the Games enhance performances?

4 Exploratory Data Analysis

- As suggested by the correlation matrix, hosting the Olympics appears to have a noticeable impact on the total medals won by the host nation. Let's verify it with another method.
- Most host countries (red dots) exceed their average performance (green dots).



Top Performing Sports by Nation

4 Exploratory Data Analysis

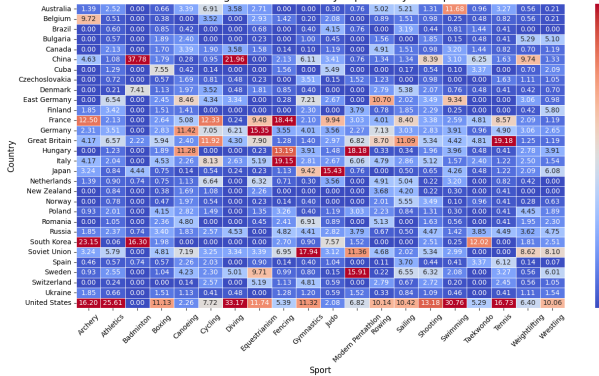


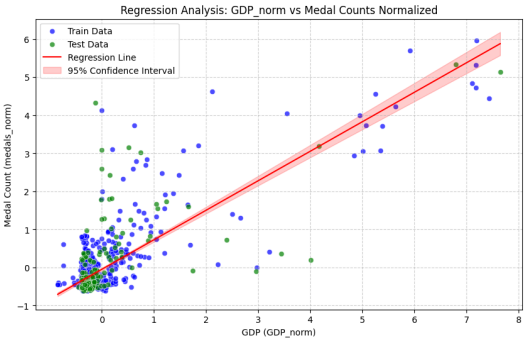
Figure: A comparison of nations' strengths in various sports, based on their medal counts across different Olympic Games.

Table of Contents

5 Simple Linear Regression

GDP_norm vs. medals_norm

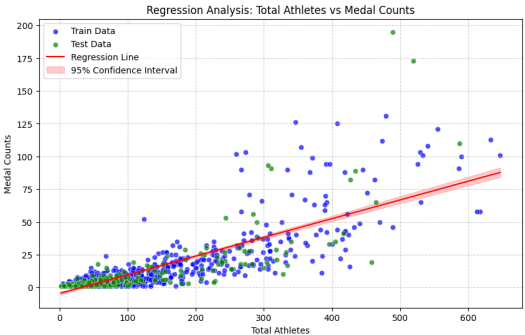
5 Simple Linear Regression



- The graph highlights a significant relationship between GDP and the number of medals obtained; In particular the $R^2 = 0.683$, meaning that more than 68% of the variance in the total medals is explained by the GDP.
- There are still some outliers, in particular there are some nations with small GDP that overperform the number of medals, examples could be small nations but with big history in specific sports.
- Notably exists a big cluster of nations with very small number of medals and GDP, including all the small countries.

Total Athletes vs. Total Medals

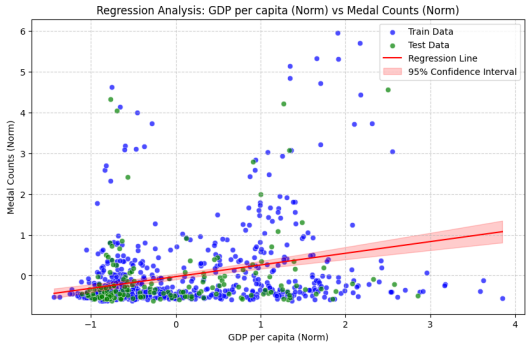
5 Simple Linear Regression



- The model captures a significant portion of the variance in medal counts (0.69%), indicating that the number of athletes is a major factor influencing medal totals.
- **Outliers**:
  - Nations with high medal counts relative to the number of athletes (overperformers). These may include small countries with strong traditions or dominance in specific sports (e.g. Jamaica in sprinting or Kenya in long-distance running).
  - Nations with a good number of athletes but relatively few medals (underperformers). It probably includes big nations with few medals.

GDP_per_capita vs. medals_norm

5 Simple Linear Regression



- Surprisingly, the relationship between GDP per capita and medals is almost null ($R^2 < 0.10$).
- This is because very small but rich countries, such as Luxembourg, San Marino, Monaco, lead in the GDP per capita ranking but obviously don't win much money because of the number of athletes they have.
- I decided to exclude this variable from my analysis due to the bias it could introduce.

Table of Contents

Multi-linear Regression

6 Multi-Linear Regression

OLS Regression Results

Dep. Variable:	medals_norm	R-squared:	0.856
Model:	OLS	Adj. R-squared:	0.855
Method:	Least Squares	F-statistic:	554.3
Date:	Fri, 13 Dec 2024	Prob (F-statistic):	1.53e-193
Time:	18:38:46	Log-Likelihood:	-208.18
No. Observations:	472	AIC:	428.4
Df Residuals:	466	BIC:	453.3
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.2668	0.119	-2.247	0.025	-0.500	-0.033
Country_Host	-0.4914	0.170	-2.893	0.004	-0.825	-0.158
gdp_norm	0.3341	0.025	13.233	0.000	0.284	0.384
Total_Athletes	0.0051	0.000	23.800	0.000	0.005	0.006
Population_perc	0.0209	0.005	4.008	0.000	0.011	0.031
HDI	-0.4914	0.161	-3.060	0.002	-0.807	-0.176

- $R^2 = 0.856$: 85.6% of the variance in medals is explained by the model.
- **Significant predictors (p-value < 0.05)**:
  - GDP, Total Athletes, Population_perc, and HDI
- `Total_Athletes` has the strongest effect (because of its very high t-statistics), followed by the GDP.
- **Coefficients interpretation:**
  - **ind. variable** (coef $= k$): A unit increase in ind.variable increases the normalized medal count by k medals, holding other factors constant.

Predictions

6 Multi-Linear Regression

- The scatterplot compares the actual medals to the predicted medals on the test set.
- The red dashed line represents perfect predictions ($y = x$). Points close to this line indicate accurate predictions.
- Most predictions align well with the line, confirming the effectiveness of the model.
- The R^2 on the test set is 0.82.
- Test MSE: 0.28.

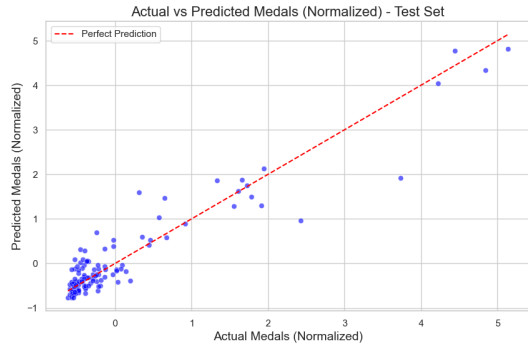


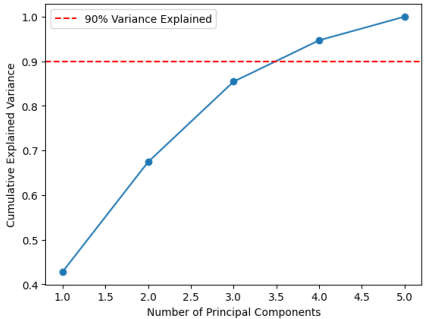
Table of Contents

7 PCA

PCA - regression analysis

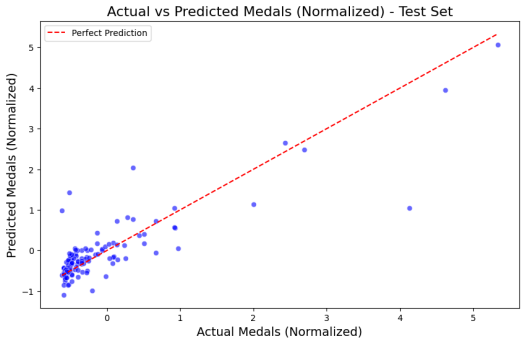
7 PCA

- **Number of PCs**:
  - **3 PCs** explain more than 85% of the variance
- **Regression Model:**
  - R^2 **(Train):** 0.77 (77% of variance explained).
  - **Cond. No.** 1.6 (Predictors are now independent).
- **Significant Principal Components:**
  - **PC1:** Strong positive impact on normalized medal count ($\beta = 0.5928$, $p < 0.001$).
  - **PC2:** Moderate positive impact ($\beta = 0.0599$, $p = 0.001$).
  - **PC3:** Significant negative impact ($\beta = -0.1743$, $p < 0.001$).



PCA - Predictions

7 PCA

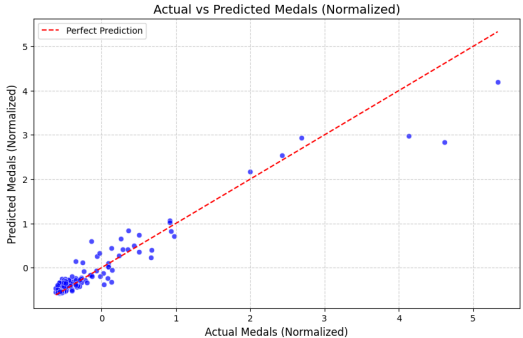


- **Prediction Metrics:**
  - **Test** R^2: 0.74
  - **Test MSE**: 0.24
- PCA was applied to address multicollinearity. However:
  - Both train and test R^2 values are noticeably lower compared to the previous model.
  - The coefficients are no longer easily interpretable, reducing the model's explanatory value.
  - The limited number of predictors does not justify the complexity introduced by PCA.
- Alternative methods will be explored...

Table of Contents

Random forest - Exceptional results

8 Random Forest



- **Performance Metrics:**
  - **Test** R^2: 0.91
  - **Test MSE**: 0.081
- **Key Insights:**
  - Outperforms linear models by handling **non-linear relationships**.
  - Captures **clusters and outliers** effectively.
  - The scatterplot shows that most points align well with the perfect prediction line.
- **Feature Importance:**
  1. **Total Athletes**
  2. **GDP**
  3. **Population**

Table of Contents

Conclusions and future works

9 Conclusions

Conclusions:

- **Economic and social factors**, such as GDP, population, and total athletes, are **significant predictors** of Olympic performance.
- **Random Forest** outperformed linear models with a test $R^2 = 0.91$, effectively capturing non-linear relationships and outliers.
- Feature importance analysis revealed that *Total Athletes*, *GDP*, and *Population* are the most influential predictors.

Future works:

- Incorporate **additional predictors**, such as investments in sports infrastructure in medal-winning countries.
- Apply **clustering techniques** (K-Means or clustering per continent) to group countries with similar Olympic performance profiles.
- Explore other ensemble methods or deep learning models to further improve predictive accuracy.

How do economic and social factors influence national performance in the Olympic Games?

Thank you for listening!

Any questions?