CSE 519 - Progress Report - The Olympic Team Challenge

Introduction

The primary objective of this project is to conduct a thorough analysis of the factors influencing a country's participation in the Olympics. The investigation delves into the trends of Olympic participation, drawing data from diverse sources such as the '120 years of Olympic history: athletes and results' dataset on Kaggle, World Bank indicators, and supplementary country information from the Global Country Information Dataset.

Data

The dataset was meticulously constructed by amalgamating information from multiple sources. The '120 years of Olympic history: athletes and results' dataset covers the period from 1896 to 2016, while the World Bank indicators dataset spans from 1960 to 2021. The World Bank dataset encompasses a spectrum of indicators spanning population, education, infrastructure, and economy.

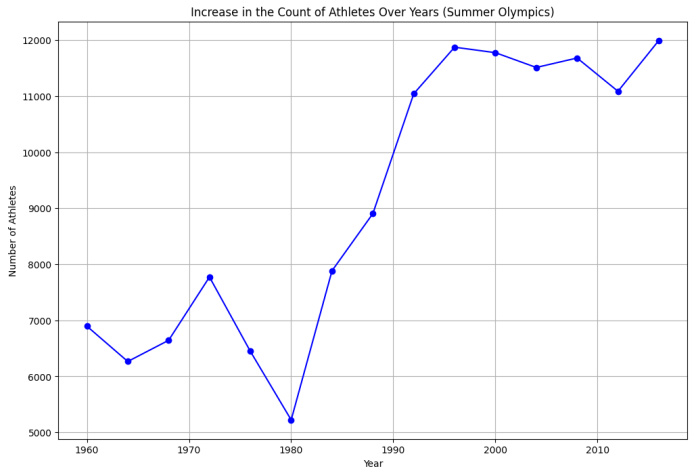
Inevitably, the dataset presented challenges, including missing values and non-standardized data. Mitigating these challenges involved the following steps:

1. **Standardization**: Before merging the datasets, a comprehensive standardization process was executed. This involved a meticulous review of the dataset, ensuring uniformity across columns.
2. **Handling Missing Values**: An algorithm was devised to calculate the average difference in values between successive rows within a specific column. This information was then utilized to compute missing values through backfill or forward-fill. Importantly, this operation was conducted on a per-country basis, acknowledging the uniqueness of individual country trends.
3. **Merge Strategy**: To facilitate integration with the Olympics dataset, the season column was added to the indicators dataset. The integration strategy employed an 'inner join' using Pandas merge on 'Country,' 'Year,' and 'Season' columns.

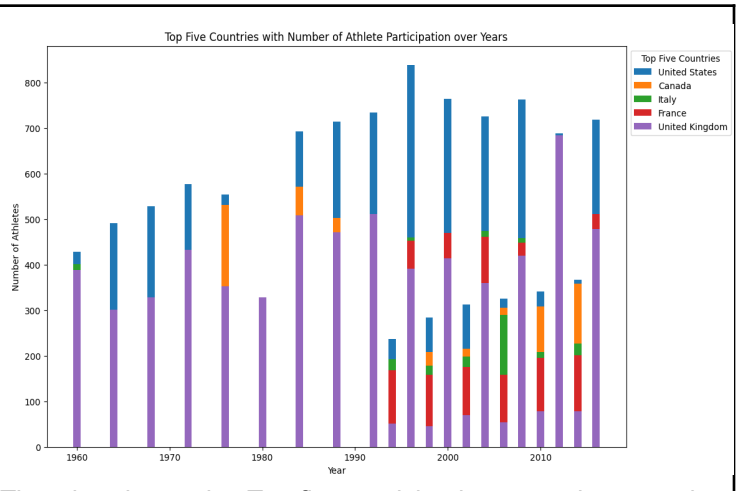
The resulting dataset was more meaningful to allow it to be usable for generating plots and insights.

Plots and Insights

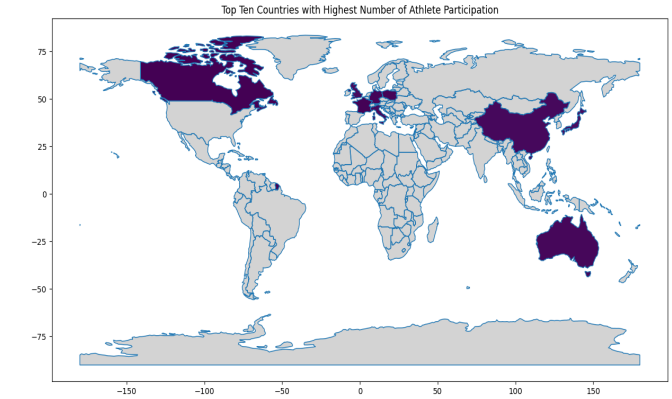
The resulting dataset presented us with various insights and plots that helped us identify the factors that are responsible for a country’s participation in the Olympics.



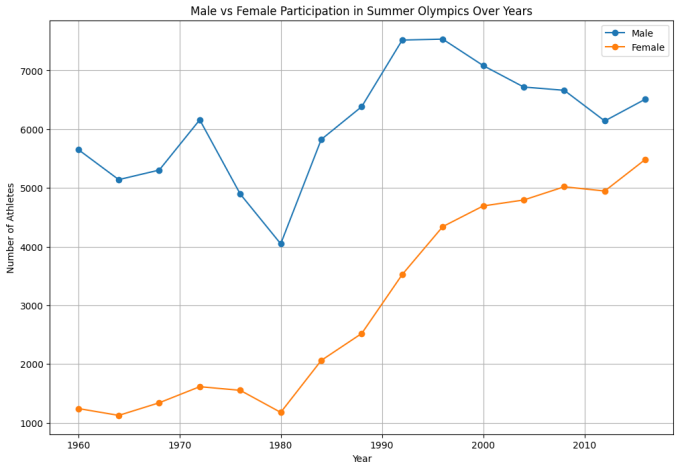
The plot shows the Trend of increment of participation in recent years. The sharp drop in 1980 was due to an Olympics boycott led by the United States and several other Western countries in response to the Soviet Union's invasion of Afghanistan in December 1979.



The plot shows the Top five participating countries over the years. The United States has been the top participant country with the most number of athletes every year except 1980.

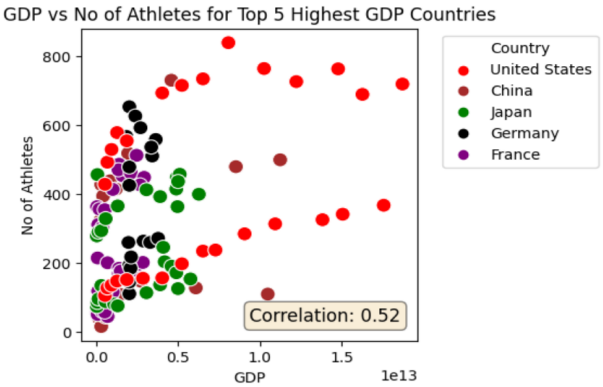


The highlighted areas in the map show the top ten participating countries in the world.



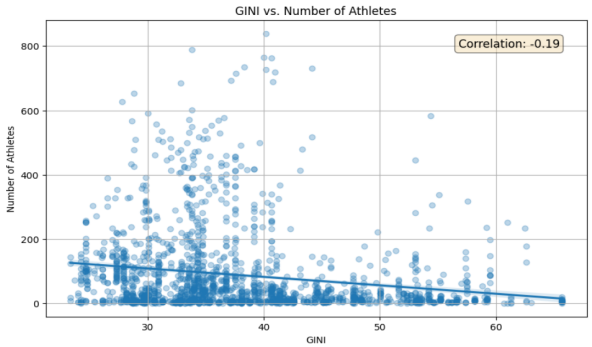
Women's participation has significantly increased after 1980.

1. GDP



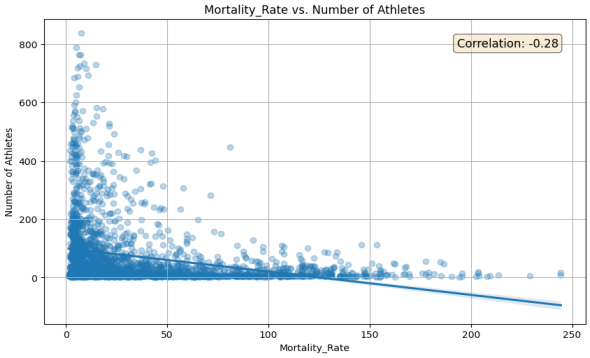
A nation's GDP significantly influences its participation in the Olympics, reflecting economic strength and resources. Higher GDP correlates with increased investment in sports infrastructure, athlete training, and the overall development of a healthier, more athletic population. The plot shows a high positive correlation between GDP and the Number of Athletes participating in the Olympics for the Top Five Countries with the highest GDP. The USA has the highest GDP and correspondingly has the highest participation among other countries.

2. GINI



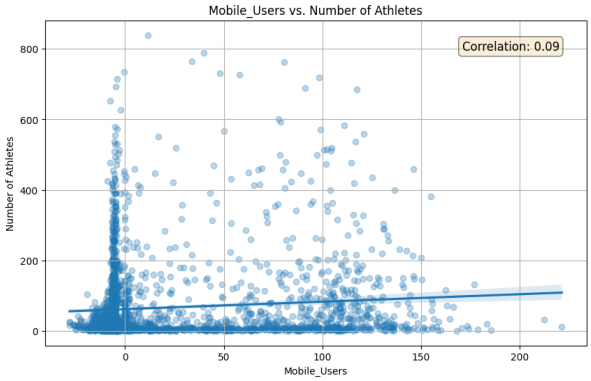
The GINI coefficient, a measure of income inequality, plays a role in a country's engagement in the Olympics. Elevated income inequality could restrict resources for sports development, influencing athletic talent. Conversely, a fairer income distribution may promote inclusivity, cultivating a diverse athlete pool and bolstering a nation's global competitiveness in sports. The plot shows GINI having a negative correlation with the number of athletes and shows Countries like Bolivia having GINI above 70 have very little participation.

3. Mortality Rate



The connection between a country's mortality rate and its participation in the Olympics is notable. Elevated mortality rates may signal health obstacles, potentially impacting the fitness and performance of the population as shown in the plot with negative correlation. Conversely, nations with lower mortality rates are likely to field stronger and more competitive contingents in the Olympic games, highlighting the relationship between health indicators and athletic success.

4. Mobile Users



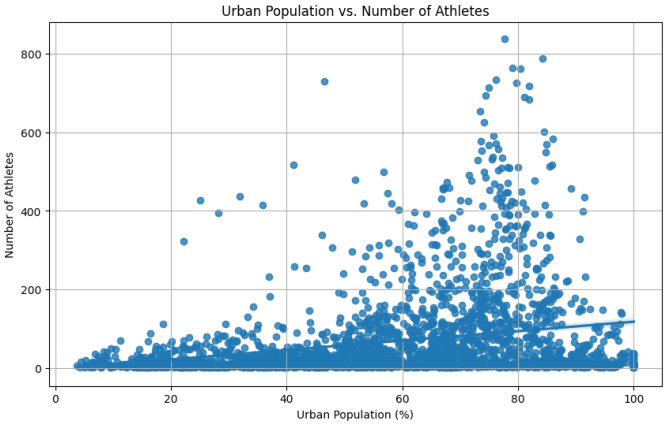
Surprisingly, the number of mobile users has no significant impact on the participation of Athletes in the Olympics. Mobile Penetration should give global connectivity for athletes and motivate them with the contents of sports, and real-time training that should have led to more participation in the Olympic games but we can see not much of a positive correlation.

Hypothesis Testing

We test five different hypotheses that could provide better insights into the factors that affect the participation or performance of countries in the Olympics:

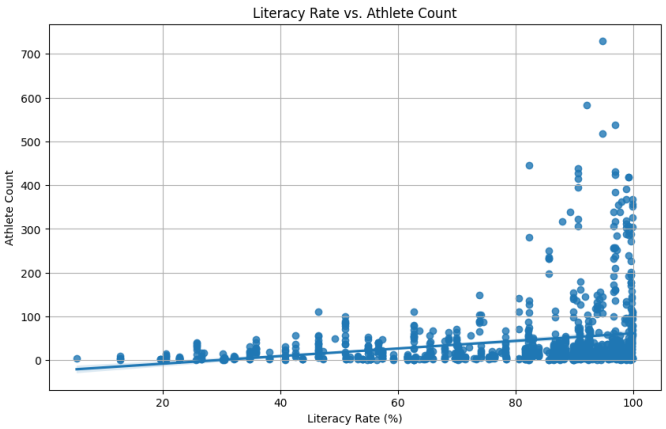
1. Relationship Between Urban Population and Olympic Participation.
2. Impact of Literacy Rate on Olympic Participation.
3. Impact of Gender Ratio on Olympic Success.
4. Relationship between Population Density and Olympic Success.
5. Effect of Political Stability on Olympic Participation.

1. Relationship Between Urban Population and Olympic Participation



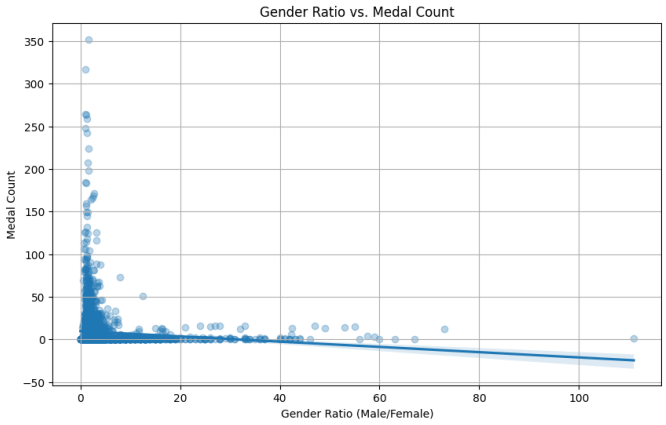
Null Hypothesis (H0): There is no correlation between the urban population percentage of a country and the number of athletes it sends to the Olympics.
Alternate Hypothesis (H1): There is a significant correlation between the urban population percentage of a country and the number of athletes it sends to the Olympics.
Pearson Correlation Coefficient: 0.259
P-value: 1.52×10^{-40}
The positive correlation coefficient suggests a weak to moderate positive relationship between the percentage of the urban population in a country and the number of athletes it sends to the Olympics. The extremely low p-value indicates that this result is statistically significant, thus we can reject the null hypothesis (H0) that there is no correlation.

2. Impact of Literacy Rate on Olympic Participation



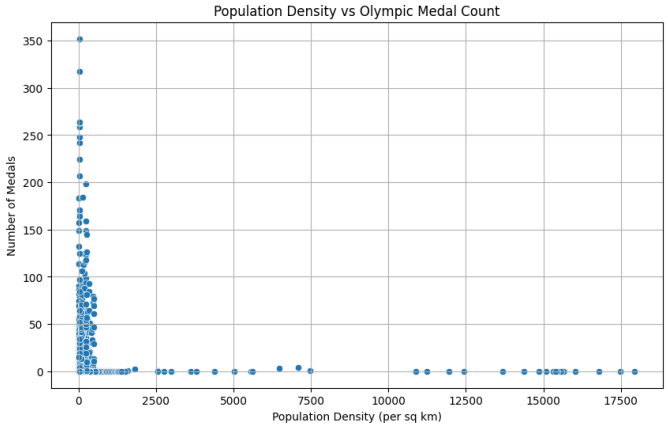
Null Hypothesis (H0): The literacy rate of a country does not significantly affect its participation in the Olympics.
Alternate Hypothesis (H1): Higher literacy rates in a country are associated with higher participation in the Olympics.
Pearson Correlation Coefficient: 0.229
P-value: 3.02×10^{-14}
This result indicates a weak positive correlation between a country's literacy rate and its participation in the Olympics. The p-value, being significantly low, suggests that this correlation is statistically significant. Therefore, we can reject the null hypothesis (H0) that states the literacy rate of a country does not significantly affect its participation in the Olympics.

3. Impact of Gender Ratio on Olympic Success



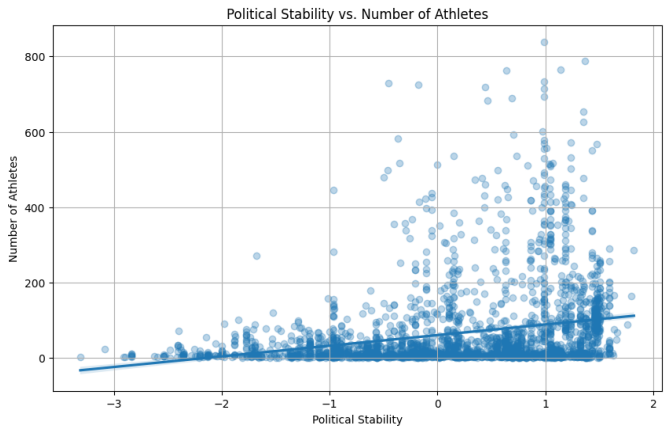
Null Hypothesis (H0): The gender ratio (male to female) of a country's Olympic team does not influence its success in terms of medal count.
Alternate Hypothesis (H1): A more balanced gender ratio in a country's Olympic team is associated with greater Olympic success.
Pearson Correlation Coefficient: −0.088
P-value: 7.55×10−6
The negative correlation coefficient indicates a weak inverse relationship between the gender ratio (male to female) in a country's Olympic team and its success as measured by medal count. The low p-value suggests that this correlation is statistically significant, allowing us to reject the null hypothesis (H0) which states that the gender ratio does not influence Olympic success.

4. Relationship between Population Density and Olympic Success



Null Hypothesis (H0): The population density of a country has no impact on its overall success in the Olympics.
Alternate Hypothesis (H1): Countries with higher population densities tend to have more success in the Olympics (measured in terms of medals won).
Pearson Correlation Coefficient: −0.038
P-value: 0.053
The correlation coefficient indicates a very weak, almost negligible negative relationship between population density and the number of Olympic medals won. The p-value is marginally above the conventional threshold of 0.05, suggesting that the result is not statistically significant. Therefore, we do not find sufficient evidence to reject the null hypothesis; thus, it appears that population density does not have a significant impact on Olympic success.

5. Effect of Political Stability on Olympic Participation



Null Hypothesis (H0): Political stability in a country does not affect the number of athletes it sends to the Olympics.
Alternate Hypothesis (H1): Countries with higher political stability send more athletes to the Olympics.
Pearson Correlation Coefficient: 0.24
P-value: 9.72×10−37
This result indicates a weak positive correlation between a country's political stability and its participation in the Olympics. The p-value, being significantly low, suggests that this correlation is statistically significant. Therefore, we can reject the null hypothesis (H0) that states the political stability of a country does not significantly affect its participation in the Olympics.

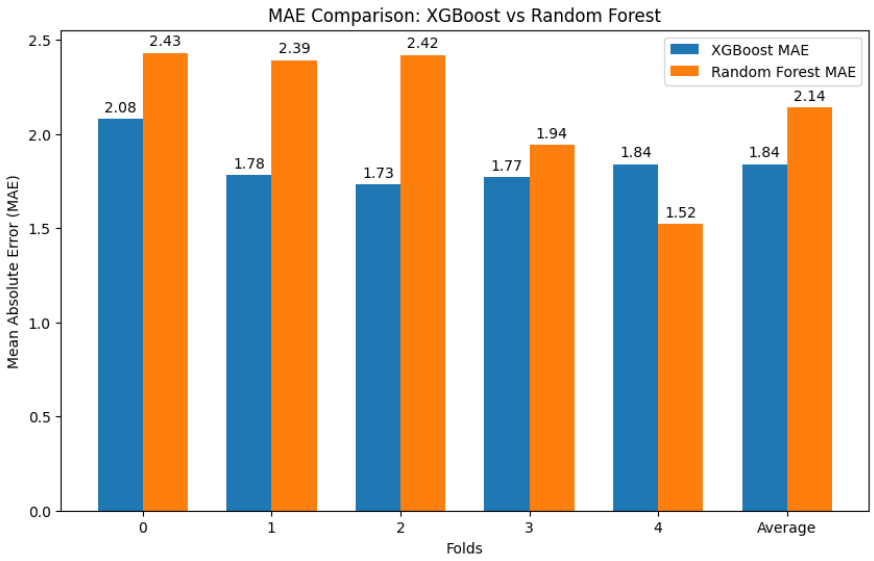
Baseline Modeling

For the Baseline Modeling, we used RandomForest and XGBoost to forecast the Countries' participation in the Olympics. We chose the number of athletes participating in the Olympics as the target variable to determine the Countries' participation.

1. Feature Engineering

a. New features were generated to train the two models. (Population Ratio, GDP Per Athlete, log of number of athletes, etc)
b. Normalization of numerical features was performed.
c. One hot encoding was used for categorical features. (Country and Season)

Five-fold validation was applied to the models for validation. The Mean Absolute Error (MAE) of the results is as shown in the diagram below. This was our Baseline model and the MAE error rate is high. This provides a room for improvement for our future work.



Timeline Progress

5/10 - 10/26 (Completed)

1. Data collection: Work on collecting and aggregating the data from the aforementioned sources, to create one source of data, consisting of indicators for each country across a range of years.

10/27 - 11/16 (Completed)

1. Data processing: Preprocessing the data including filling missing values, converting into different formats, etc.
2. Generating insights: Create charts that plot the Olympic participation trends with each of the indicators, across different countries. Generate insights based on the charts.
3. Hypothesis testing: Propose a few hypotheses on what it takes to have good participation and win, and perform statistical analysis and hypothesis testing to prove or disprove them.

11/17 - 12/7 (In Progress)

1. Modeling: Build machine learning models on the data to predict the countries' participation in the Olympics.
A significant amount of work needs to be done for feature cleaning and selection, hyperparameter tuning, and testing other machine learning models to achieve a higher accuracy.
2. Prediction: Predict the number of teams and winners for the 2024 Summer Olympics to be held in Paris.
3. Documentation: We additionally commit to publishing our work through various means such as a public GitHub repository for code, a public Kaggle dataset that hosts the data we have collected, and a blog post with our insights.