Rohan Giri, Jennifer Gonzalez,
Sameer Khan, Kristen Lowe,
María Laura Peña, Scott Stempak

Predicting a country's Olympic success

Presented to:
Deepayan Chakrabarti
Red McCombs School of Business
The University of Texas at Austin
2110 Speedway
Austin, TX 78705

August 12, 2024

Introduction

In this project, we investigate the relationship between a country's resources and its performance at the Olympic Games. The analysis merges four comprehensive datasets: the Olympic Summer & Winter Games data from 1896 to 2022, Country Population data from 1960 to 2022, World GDP data from 1960 to 2022, and additional country-specific resources such as area and coastline ratio. By combining these datasets, we aim to answer the question: "To what extent does a country’s resources impact its performance at the Olympics?" This exploration considers factors such as population, GDP, geographical characteristics, and whether the country is a host nation, and examines how these variables influence medal counts across different Olympic seasons.

Problem Importance

The importance of this problem lies in understanding the factors that contribute to Olympic success, which is a matter of national pride and global recognition. By analyzing how resources like GDP, population, and geographical characteristics influence a country's performance, this study sheds light on the broader socio-economic and geopolitical factors that drive success in international sports. This analysis is particularly relevant to policymakers, sports federations, and economists, as it could inform resource allocation and strategic planning for future Olympic participation. Moreover, the findings could resonate with a large audience, including sports enthusiasts and analysts, who are interested in the dynamics behind Olympic achievements and the potential economic impact of investing in sports infrastructure and talent development.

Methodology

To prepare the data for analysis, we took several systematic steps to ensure its accuracy and relevance. First, we collected data from multiple sources, including Olympic Games records, World GDP figures, and country population statistics from 1960 to 2022. We then cleaned the data by mapping country names to ensure consistency across datasets, pivoted data frames to align with our analytical needs, and removed any missing values that could skew results. Once the datasets were ready, we merged them. Significant efforts were made in feature engineering, where we derived variables such as GDP per capita, coastline-to-area ratio, and made host country status a binary variable. These steps were crucial in creating a cohesive dataset that allowed

for a robust examination of how various country resources impact Olympic performance.

Exploratory Data Analysis

After cleaning and wrangling, we ran descriptive statistics to get a broad look at trends and patterns within our data. One initial interesting observation was that between 1960 and 2020, the number of countries who won at least one Olympic medal increased by nearly 200%; this may be because of reasons like increased globalization and resource allocation towards sports as countries developed economically, or simply because of the emergence of new countries.

There was variability across metrics with GDPs ranging from $222 million to over $21 trillion, and populations ranging from 23,000 to 1.4 billion, reflecting the diverse backgrounds of participating nations. Additionally, most countries in the dataset were not hosting the Games, suggesting that many nations compete without the advantage of being a host country. In terms of the target variable, it was found that Western Europe leads with the highest overall medal count, more than doubling second-place Asia's total count (**see Appendix A**). Countries such as the USA, Russian Federation, and Germany have performed well in both seasons (Summer and Winter) with the USA having the highest count of medals overall, however the top ten rankings change according to the season (**see Appendix B**). Further, countries with larger populations tend to win more medals with Western Europe and North America being notable high-population, high-medal regions, while countries like Oceania and The Baltics represent low-population, low-medal regions; this positive trend is true for the variable of GDP as well (**see Appendix C and D**). Seasonally, countries tend to win more medals in the Summer than in the Winter, which could reflect the greater number of events and participants in the Summer (**see Appendix E**). One more interesting observation is that being the host country positively impacts the amount of medals won, likely due to factors including increased investment in athletes and infrastructure and home advantage (**see Appendix F**).

Despite these observations, our correlation matrix resulted in very weak correlations; with the only moderate positive correlations coming from a country’s GDP and medal counts. Other slightly positive correlations included

country population and area (**see Appendix G**). This is to say that while economic strength is important, other factors like long-term planning, cultural emphasis, and targeted investments may be more critical for Olympic success.

Solutions and Insights

We hypothesized that multiple factors contribute to a country's success in winning medals at the Olympics. For our regression analysis, we selected features that encapsulate various aspects of a country's profile: host country status (a binary variable indicating whether the country is hosting the Olympics), population, region, area (in square kilometers), coastline (as a coast-to-area ratio), GDP, GDP per capita, and population density. These predictors were chosen to represent economic strength (GDP and GDP per capita), regional trends (region), geographical factors (area, coastline), and demographic characteristics (population and population density). We also sought to examine the potential "home-field advantage" by including host country status.

Given the significant differences between the medal counts at the Winter and Summer Olympics, we split the dataset by season and trained models separately for each. We employed four types of models: a base decision tree classifier, a bagging model, a random forest, and a gradient boosting model. After initial testing, we performed a grid search to optimize the hyperparameters for the bagging model, random forest, and gradient boosting model.

For each model, we evaluated performance using several metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2. For the Summer Olympics models, the optimized gradient boosting model performed best on the test set, with MAE of 4.539, MSE of 47.197, RMSE of 6.870, and R^2 of 0.858 (**see Appendix H**). For the Winter Olympics models, the optimized random forest performed best, with MAE of 3.616, MSE of 27.948, RMSE of 5.287, and R^2 of 0.704 (**see Appendix I**).

In both the optimized gradient boosting model for the Summer Olympics and the optimized random forest for the Winter Olympics, GDP and population emerged as the most important features (**See Appendix J and K**). In contrast, the region a country belongs to, such as the Baltics, did not

significantly contribute to predicting a country's medal count. These findings align with our expectations, as economic and demographic factors like GDP and population directly influence the resources available for Olympic training and the potential talent pool from which a country can select its best athletes.

Conclusion

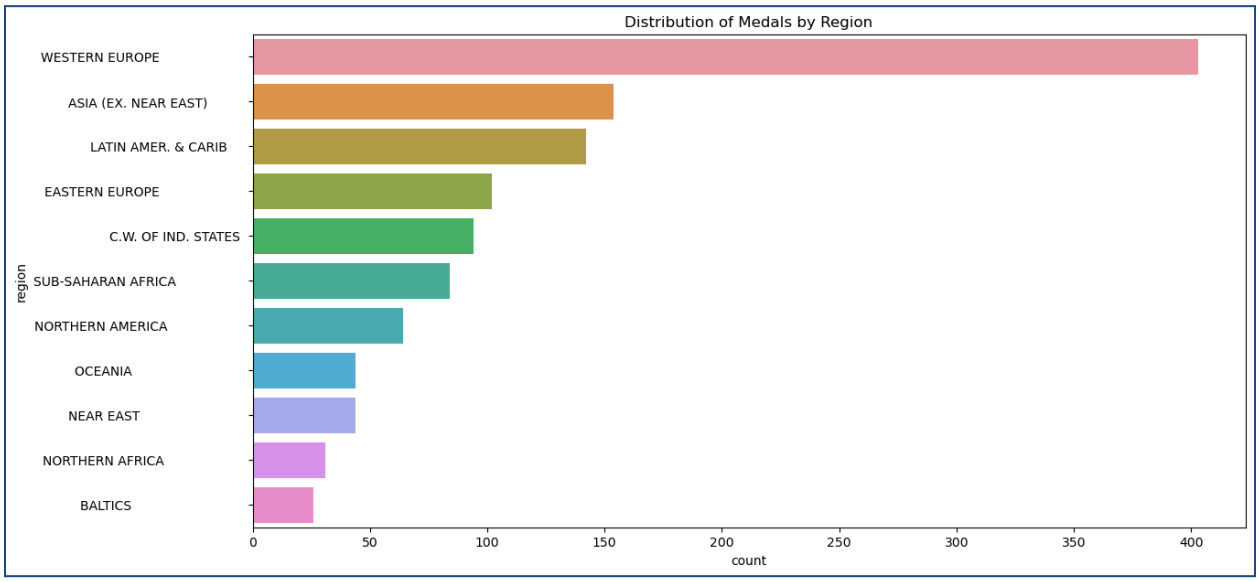
After making the models, we used the highest performing model for the Summer Olympic games (optimized gradient boosting) to attempt to predict how many medals the United States will win in the 2024 games. Our model predicts that the United States will win 114 models this year.

Overall, we found that GDP and population are the most important predictors in determining a country's success in winning medals at the Olympics. However, there are a number of limitations to this project. First of all, we were not able to take into account a country's cultural emphasis on sports, since it is not a quantifiable metric. Moreover, we didn't use any other economic indicators as predictors besides GDP. Given more time, we would have liked to include other predictors such as literacy rate, infrastructure, unemployment rates, and many others.

Appendices

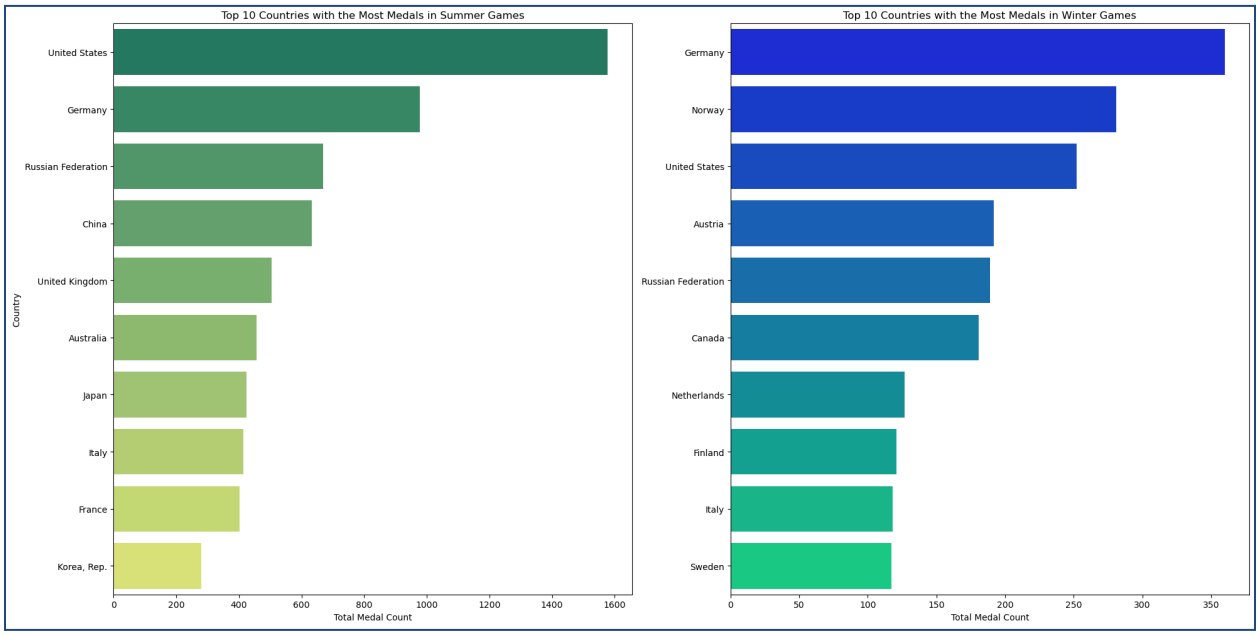
Appendix A

Distribution of Olympic Medals by Region



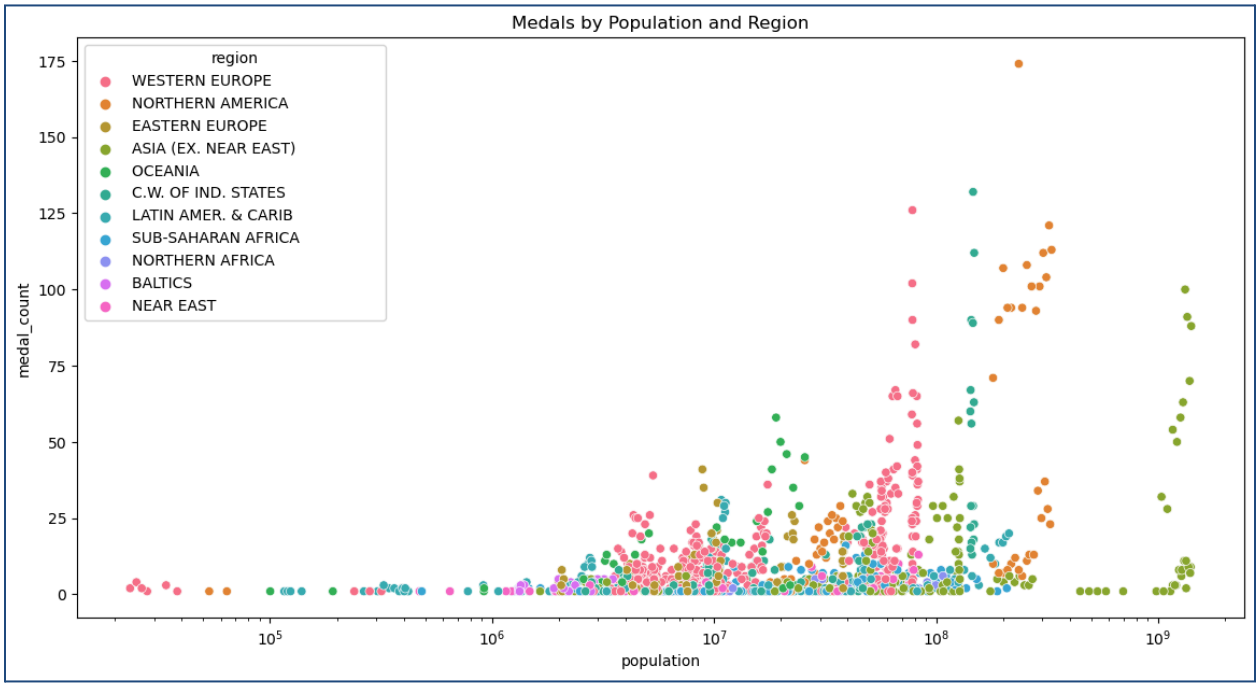
Appendix B

Top 10 Countries with the Most Olympic Medals by Season



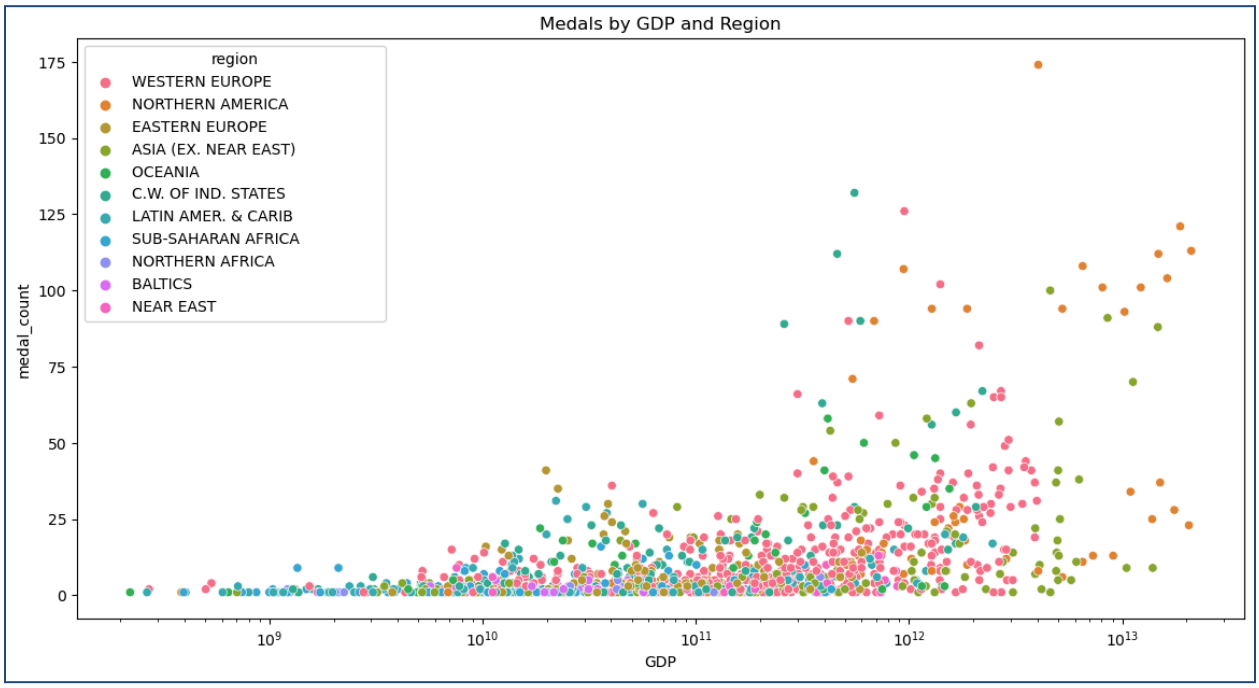
Appendix C

Distribution of Olympic Medals by Population and Region



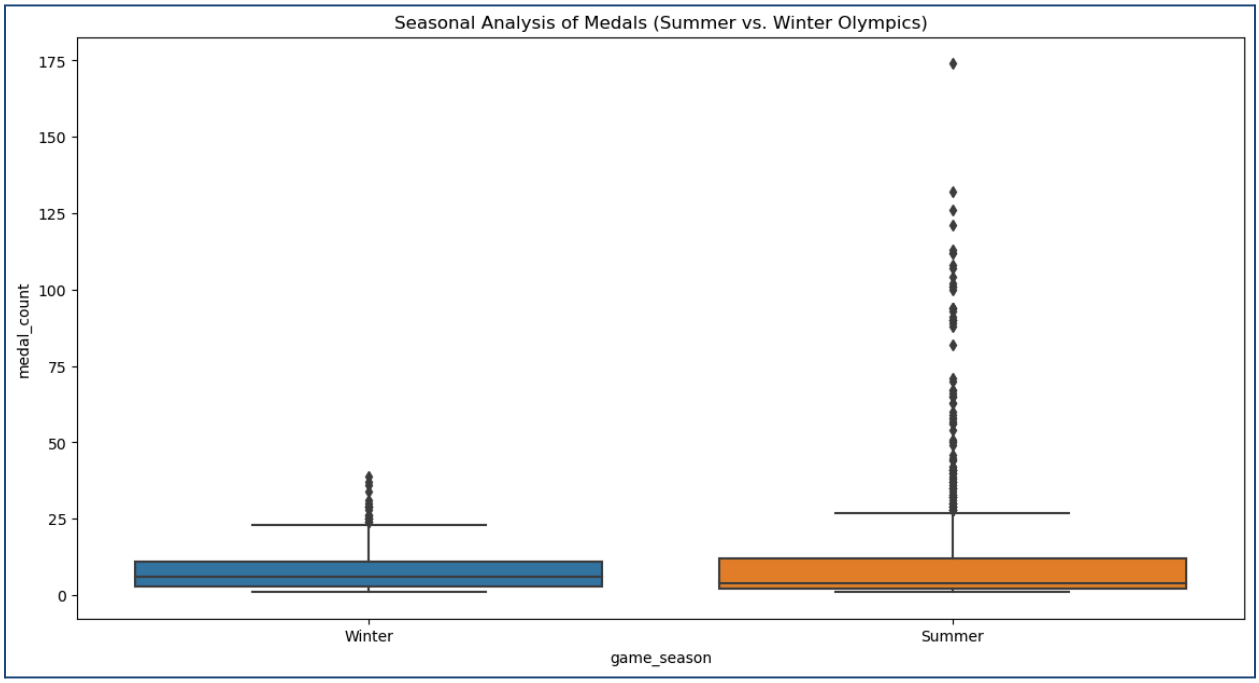
Appendix D

Distribution of Olympic Medals by GDP and Region



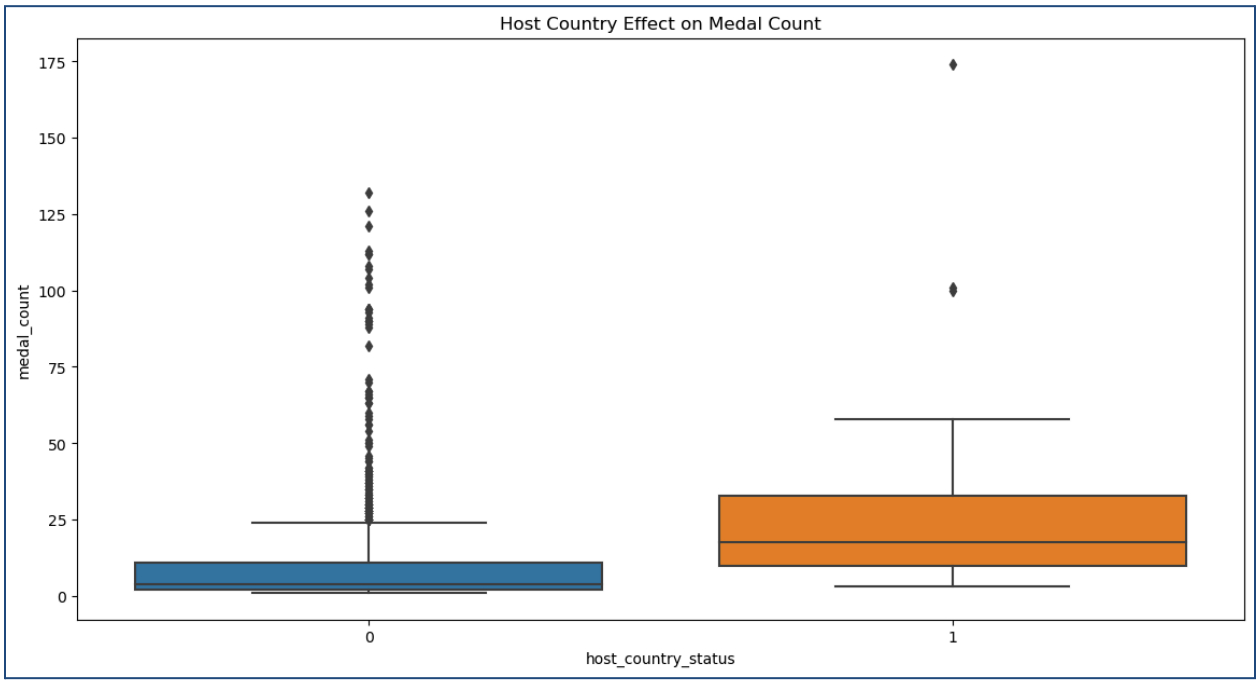
Appendix E

Seasonal Analysis of Olympic Medals



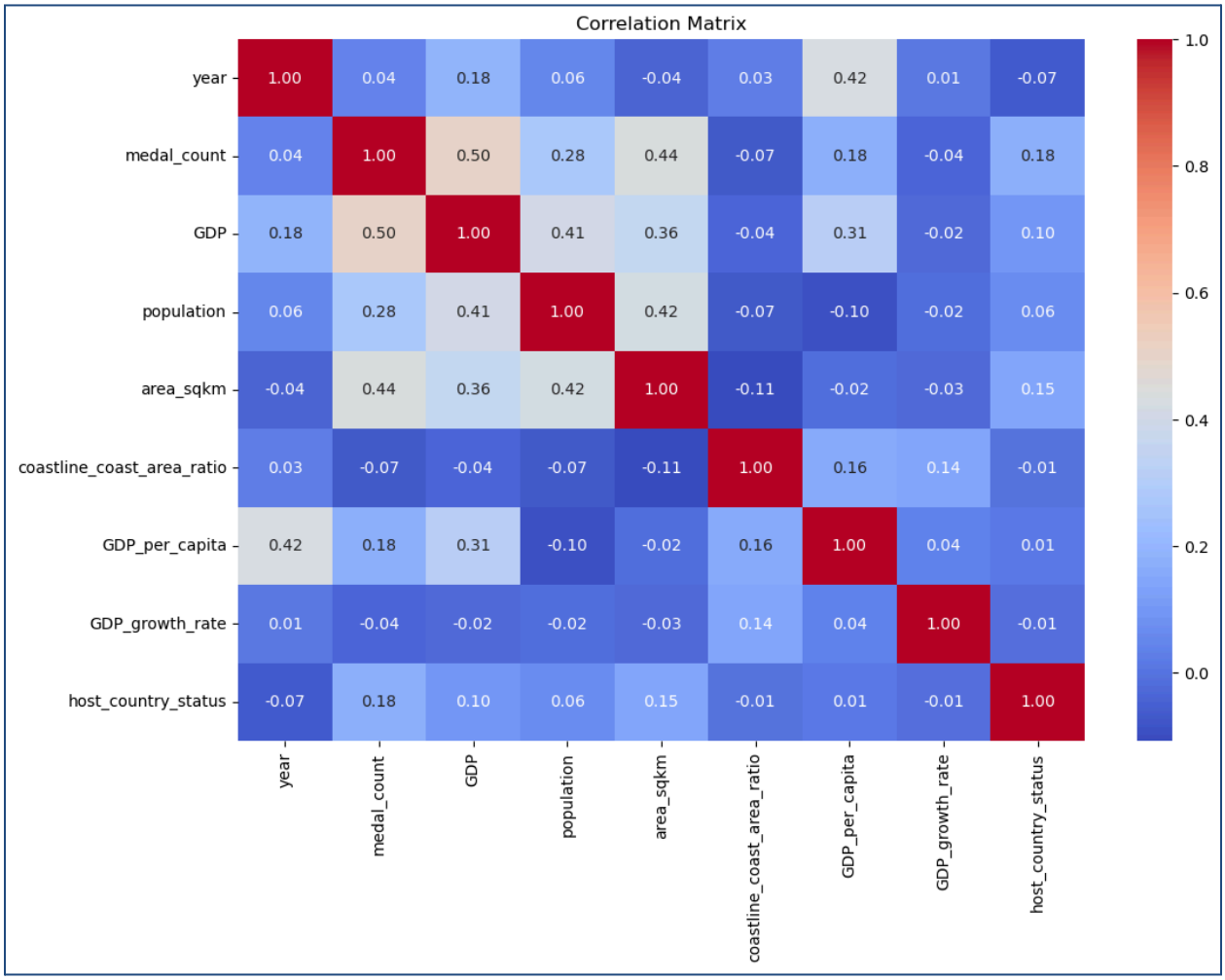
Appendix F

Host Country Effect on Medal Count



Appendix G

Correlation Matrix for Numerical Variables



Appendix H

Summer Olympic Model Evaluation Metrics

	MAE	MSE	RMSE	R^2
Decision Tree	7.786	224.552	14.985	0.323
Bagging	5.202	85.333	9.238	0.743
Random Forest	4.674	52.975	7.278	0.840
Gradient Boosting	5.014	66.179	8.135	0.800
Optimized Bagging	4.794	57.540	7.586	0.826
Optimized Random Forest	4.507	49.951	7.068	0.849
Optimized Gradient Boosting	4.539	47.197	6.870	0.858

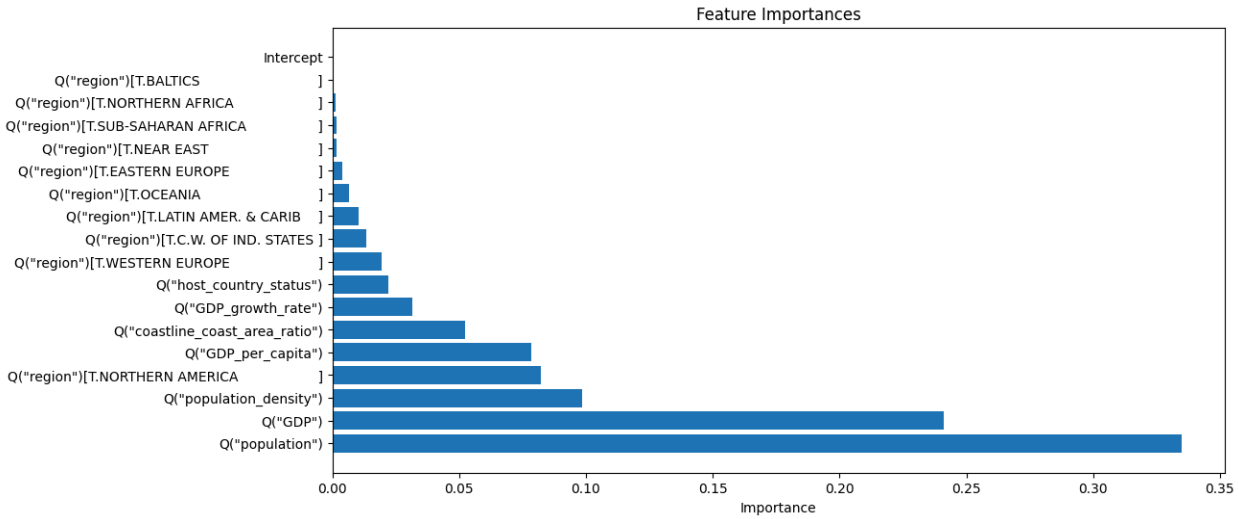
Appendix I

Winter Olympic Model Evaluation Metrics

	MAE	MSE	RMSE	R^2
Decision Tree	5.224	49.397	7.028	0.485
Bagging	3.950	30.018	5.479	0.687
Random Forest	4.060	31.426	5.606	0.672
Gradient Boosting	4.152	32.578	5.708	0.660
Optimized Bagging	3.884	30.226	5.498	0.684
Optimized Random Forest	3.616	27.948	5.287	0.709
Optimized Gradient Boosting	3.675	28.368	5.326	0.704

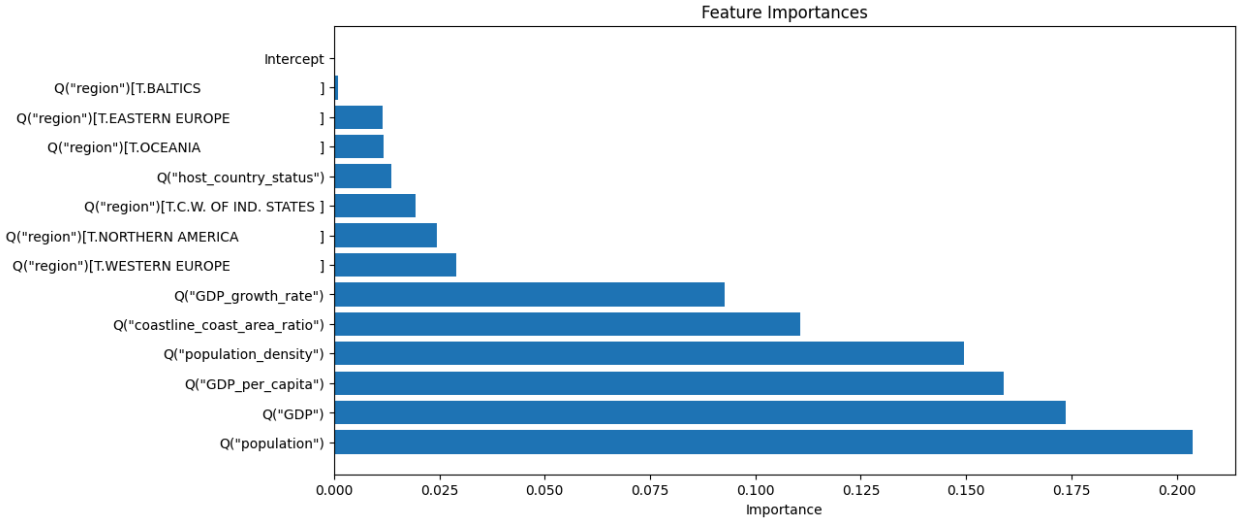
Appendix J

Summer Olympic Feature Importances



Appendix K

Winter Olympic Feature Importances



References

https://www.kaggle.com/datasets/piterfm/olympic-games-medals-19862018/data?select=olympic_results.csv

https://www.kaggle.com/datasets/annafabris/world-gdp-by-country-1960-2022

https://www.kaggle.com/datasets/ayushparwal2026/country-population-from-1960-to-2022/data

https://www.kaggle.com/datasets/fernandol/countries-of-the-world/data

https://now.tufts.edu/2024/07/17/why-do-some-countries-win-more-olympic-medals