CSE 519 - Project Proposal - The Olympic Team Challenge

Introduction

The spirit of competition between the countries has converged on the global stage through the Olympic Games. In this project, we seek to explore the trends in countries' participation in the Olympics by studying various indicators in the underlying data.

Literature review

In recent years, there has been a shift from measuring a country's progress solely based on financial indicators like GDP to assessing the happiness of its citizens. Research, such as the Harvard Study of Adult Development [1], has emphasized the importance of relationships and social connections in people's happiness. The World Value Survey, conducted in collaboration with the European Values Study, spanning from 1981 to 2020 in 102 countries [2], provides insights into what individuals desire and believe in. Researchers [3] have found that the excess mortality burden on men is influenced by a combination of biological, behavioral, and social factors, with happiness potentially contributing to longer life through factors like health, financial well-being, and social connections.

There has been a wide range of studies for predicting the factors that relate to the participation and medal count of countries in the Olympics. The size of the economy is strongly linked to a country's performance at the Olympics [6], providing abundant resources [7] and thus receiving more medals [8]. W. SHASHA ET AL show the relationship between economic, demographic, geographic, and social factors and sports success, focusing on the Rio 2016 Olympics. The findings suggest that inflation rate, economically active population, income classification, and temperature influence a country's medal ranking performance, while GDP size, corruption ranking, number of athletes, and topography do not have a significant impact [5]. Research also shows that the larger the population, the more gifted players with physical attributes are more likely to participate. Different geographical environments spawned various sports. Varied sports events are formed by topography, mountains, rivers, and different geographical environments. In their 2013 study, Feizabadi and colleagues [9] examined the correlation between a country's performance in the 2010 Guangzhou Summer Asian Games and various demographic and economic factors. They employed statistical methods including the Kolmogorov–Smirnov test, one-way Analysis of Variance (ANOVA), and Stepwise Multiple Regression (SMR) Analysis. The results indicated a noteworthy connection between a nation's success in the Guangzhou 2010 Asian Games and all demographic and economic factors considered, such as population, GDP, health expenditure, growth rate, team size, and former hosting experience.In 2019, Rosas and Flegl [10] explored the influence of GDP, corruption, and various social factors on the performance of countries in the Rio 2016 Olympic Games. They employed the Ordinary Least Square (OLS) method and identified that a country's income classification based on gross national income, the size of its economically active population, and levels of corruption were linked to enhanced performance in the Olympic Games, while inflation did not show a significant impact.

Data

It is rare to come across multiple sources of data for Olympic participation, but one reliable source of data is the '120 years of Olympic history: athletes and results' from Kaggle: https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results

The range of data is impressive, ranging from 1896 to 2016.

The World Bank is an international financial institution that provides loans and grants to governments across the world to pursue capital projects. They are also the largest source for world country data across the range of years. The data itself is categorized into several indicators. The list of all indicators can be found here: https://data.worldbank.org/indicator. The range of years the data is available is usually from 1960 to 2021.

The following is the list of hand-picked indicators that we believe best represent a country in general and in sports (listed categorically):

Population

1. Urban population (% of total population): https://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS
2. Population ages 0-14 (% of total population): https://data.worldbank.org/indicator/SP.POP.0014.TO.ZS
3. Population ages 15-64 (% of total population): https://data.worldbank.org/indicator/SP.POP.1564.TO.ZS
4. Net migration: https://data.worldbank.org/indicator/SM.POP.NETM
5. Mortality rate, infant (per 1,000 live births): https://data.worldbank.org/indicator/SP.DYN.IMRT.IN
6. Fertility rate, total (births per woman): https://data.worldbank.org/indicator/SP.DYN.TFRT.IN
7. Population density (people per sq. km of land area): https://data.worldbank.org/indicator/EN.POP.DNST
8. Population, female (% of total population): https://data.worldbank.org/indicator/SP.POP.TOTL.FE.ZS

Education

1. Primary completion rate, total (% of relevant age group): https://data.worldbank.org/indicator/SE.PRM.CMPT.ZS
2. Literacy rate, adult total (% of people ages 15 and above): https://data.worldbank.org/indicator/SE.ADT.LITR.ZS

Infrastructure

1. Access to electricity (% of population): https://data.worldbank.org/indicator/EG.ELC.ACCS.ZS
2. Mobile cellular subscriptions (per 100 people): https://data.worldbank.org/indicator/IT.CEL.SETS.P2
3. Land area (sq. km): https://data.worldbank.org/indicator/AG.LND.TOTL.K2

Economy

1. GDP (current US$): https://data.worldbank.org/indicator/NY.GDP.MKTP.CD
2. GDP per capita (current US$): https://data.worldbank.org/indicator/NY.GDP.PCAP.CD
3. Foreign direct investment, net inflows (BoP, current US$): https://data.worldbank.org/indicator/BX.KLT.DINV.CD.WD
4. Inflation, consumer prices (annual %): https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG
5. Gini index: https://data.worldbank.org/indicator/SI.POV.GINI
6. Military expenditure (% of GDP): https://data.worldbank.org/indicator/MS.MIL.XPND.GD.ZS
7. Unemployment, total (% of total labor force) (modeled ILO estimate): https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS
8. Labor force, total: https://data.worldbank.org/indicator/SL.TLF.TOTL.IN

9. Labor force, female (% of total labor force): https://data.worldbank.org/indicator/SL.TLF.TOTL.FE.ZS

For general information about countries such as their location (latitude and longitude), official language, etc, we would be using the Global Country Information Dataset 2023 from Kaggle: https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023

Proposal

Goals for this project:

1. Prepare a dataset for Country-wise participation in the Olympics. Combine the Olympic dataset with general data about the countries.
2. Analyze and generate insights on what factors contribute to participation in the Olympics. Also, test various hypotheses using the data.
3. Use machine learning algorithms to model the data and construct predictive models.
4. Predict the number of teams/athletes that would be sent from different countries given the past data and current indicators.

In this project, we propose to start off with exploratory data analysis and hypothesis testing to better understand the factors responsible for the participation of various countries in the Olympics. Then, we plan to use those insights to model the data better and create predictive models based on various machine learning algorithms including but not limited to RandomForest, boosting models, neural networks, auto-regressive models, etc.

As one might have noticed, the ranges for the Olympic dataset and the World Indicators dataset do not overlap completely. While the Olympic dataset is from 1896 to 2016, the World Indicators dataset is from 1961 to 2021. This means that our range is the intersection of the two: 1961 to 2016, which gives us 56 years of data to work with, which is plenty of data to reduce the variance in the ML models.

Timeline

5/10 - 10/26

- 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

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

11/17 - 12/7

- Modeling: Create predictive models based on the data to predict participation in the Olympics.
- Prediction: Predict the number of teams and winners for the 2024 Summer Olympics to be held in Paris.

References

[1] https://www.adultdevelopmentstudy.org/
[2] https://www.worldvaluessurvey.org/wvs.jsp
[3] Kayonda Hubert Ngamaba, Ngianga-Bakwin Kandala, Francois Batuyekula Ilenda, “Perseverante Kawata Mupolo Are men’s happiness and life satisfaction linked to why men die earlier than women? A panel study from 1981 to 2020 in 102 countries” https://www.journalofhappinessandhealth.com/index.php/johah/article/view/33/15
[4] Millan, R., & Esteinou, R. (2021). Family satisfaction in Latin America: Do relationships matter? Perfiles Latinoamericanos, 29(58). doi:10.18504/pl2958-012-2021
[5] Wang Shasha, Babar Nawaz Abbasi, Ali Sohail “Assessment of Olympic performance in relation to economic, demographic, geographic, and social factors:quantile and Tobit approaches “ https://hrcak.srce.hr/file/438677
[6] Luiz, J. M., & Fadal, R. (2011). An economic analysis of sports performance in Africa.
International Journal of Social Economics, 38(10), 869–883. https://doi.org/10.1108/
03068291111170415
[7] Debroy, B. (2011). Does GDP growth influence sporting performance? https://economictimes.indiatimes.com/opinion/et-editorial/does-gdp-growth-influence-sportingperformance/articleshow/7851900.cms
[8] Sen, S. (2021). Does the economy determine a country’s performance at the Olympics? https://www.thehindu.com/data/data-does-economy-determine-a-countrys-performance-at-olympics/article35899178.ece
[9] Feizabadi, M. S., Khabiri, M., & Hamid, M. (2013). The relationship between the success of
countries at the Guangzhou 2010 summer Asian games and demo-economic factors.
Procedia - Social and Behavioral Sciences, 82(2013), 369–374. https://doi.org/10.1016/j.sbspro.2013.06.277
[10] Rosas, L. A. A., & Flegl, M. (2019). Quantitative and qualitative impact of GDP on sport performance and its relation with corruption and other social factors. Noesis. Revista de
Ciencias Sociales y Humanidades, 28(1), 15–37. https://doi.org/10.20983/noesis.2019.1.2