Why Johnny Can’t Win Medals

Ben Greenberg, Spencer Howell, and Daniel Troutman

Abstract— **For our project, we will be attempting to determine major factors in deciding how many medals each country will take home from each Olympics. We will gather data about each country such as population, GDP, etc. from sources such as the World Bank. Then we will compare this data with the medal counts from aggregated sources such as Kaggle datasets, scraped data from the official Olympics website, etc. From this data we hope to find which factors have more influence over medal counts over others. Should we have enough time after completing our analysis, we will attempt to develop a machine learning algorithm that will attempt to predict the results of the next Olympics based on current country statistics.**

I. INTRODUCTION

The Olympics has been a worldwide phenomenon since 1896, with more than 200 nations participating in both the Summer and Winter Olympic Games. The Games have a rich history dating back to the ancient Greek tradition, and it still maintains the interest of fans worldwide.

With such a storied history and a worldwide appeal, the Olympic Games provide a wealth of data to analyze, along with a wide number of interesting questions that can be answered using this data. Our team plans to dive into these datasets in order to find new insights and answer some of these questions.

II. MOTIVATION

While the modern Olympic Games are hosted by a private organization, they have always been inherently political, with nations invested in representing themselves well on the worldwide stage. Therefore, great importance has been placed on the medals won by each country at the Olympic games.

The fierce competition and intense pressure placed on the athletes has led to a large number of scandals and accusations surrounding the Games. In 2019, Russia was banned for 4 years for a state-sponsored program for performance-enhancing drugs [1]. This is the most recent in a long string of doping scandals associated with the Games. Some countries and athletes will go to extreme lengths to win medals.

We would like to investigate the various factors that might influence a country’s performance at the Olympics. In this analysis, we can analyze factors of a country’s economy, population, social life, and more throughout the years and see if and how they correlate with the medal count at the Olympics.

Through this analysis, we hope to answer questions such as:

- What factors most correlate with success at the Olympic Games?
- How fair are the Games for countries with smaller populations or less wealth?
- What countries win a disproportionate number of medals at the Olympics relative to their population size?
- What social factors can influence a country’s performance at the Olympics?

III. DATA COLLECTION PROCESS

In order to analyse Olympics data, we must first find many years of results for all the events. Even if we cannot find a dataset with all the needed information, we can aggregate multiple ones to fill in gaps. We will first search for ready to use datasets from websites like Kaggle or Google’s Dataset Search. If we cannot find any ones from the community, we can also use a web scraper to extract the information from websites. We can also restrict our scope to fewer events, years, or other data points if we have to manually collect it directly from the web. Once we have Olympics data, we can get county specific data from similar sources. This could include statistics about their economy, population, culture, or politics. This information may not be as laid out and organized as Olympics data, so we will have to rely more on scraping the internet.

After collecting all the data, we can use Python and several third party libraries such as Pandas and Numpy to parse and explore the data. We can use Plotly to graph the data and emphasize certain areas of interest. Finally, if there are connections between country statistics and their Olympics performance, we can attempt to predict outcomes with machine learning libraries like SciKit Learn and XG Boost. All of our findings will be recorded in a report that we can showcase at the end of this class.

IV. MEMBER RESPONSIBILITIES

For this project, we will try our best to make sure that the work is evenly distributed amongst everyone in the group. We will try to split up the work in a way that everyone can be working on some aspect of the project at the same time without anyone remaining too idle. For instance, work could be split up into web scraping, data cleaning, analysis, and possibly machine learning training. Responsibilities will be distributed based on workload as well as experience in each component of the project. We’ll rely on GitHub issues for progress tracking and project management due to the scale of our project.

One team member will be designated as the project manager for each sprint, and this role will rotate between our team members as the project progresses. The project manager will be responsible for dividing work among the members and ensuring that approaching deadlines are met. This will give each team member experience with management and keep our team on schedule.

V. Timeline

A. Data Collection and Organization - Sprint 1

In our first sprint, the team will focus on finding quality data sources, scraping and cleaning data where needed, and collecting them into well-formatted documents that can be used for our analysis. We will need to generate a list of all medals won by each country represented in the Olympics, grouped by year. We will also need to generate lists of global metrics to compare against these medal counts, also organized by year and country.

B. Data Analysis and Comparison - Sprint 2

In the second sprint, the team will focus on taking these data sources and conducting data analysis on the various sources to answer questions generated by ourselves and by our peers. We will use tools such as Jupyter Notebooks and Python data science libraries to conduct this analysis, as well as other libraries and tools to visualize the results.

C. Report Findings - Sprint 3

In the final required sprint, our team will collect the most convincing and interesting findings and present them in a graphical, easy-to-understand format in our final report. We will explain our process and how it led to our results, as well as provide the insights we gained throughout the project.

D. Machine Learning Development (TBD) - Sprint 4

Should we have enough time after sprint 3, our team will attempt to develop a machine learning model that will be able to predict the outcomes of future Olympic games based on the significant factors we find in prior sprints. We will use tools such as SciKit Learn and XG Boost to develop our models and we will train our models on the data we gathered before this sprint.

VI. Expected Outcome

While our team is unsure of specific results we expect to find, we have a general idea of the trends we expect to see. We expect that wealthier nations will have greater success at the Olympics than less wealthy nations, due to their increased resources to dedicate to training athletes. We expect countries with higher populations to generally perform better as well, since more citizens provides a larger talent pool to pull elite athletes from. However, we also expect to find some nations with a disproportionate number of medals compared to their population, which we can investigate the reasoning for.

Our team is interested in challenging these assumptions and letting the data show us insights and new ideas that were previously unexpected.

References

[1] J. Rathborn, “Why is Russia banned from the Olympics and what is ROC?,” MSN, Jul. 08, 2021. https://www.msn.com/en-gb/sport/olympics/why-is-russia-banned-from-the-olympics-and-what-is-roc/ar-AAMTfox (accessed Sep. 26, 2021).