Predictive Analysis of 120 Years Olympics History

Stephanie Tanasia Saputra
40199284

I. Introduction

Olympic games have been around for almost 3,000 years and their significance has yet to dim. People from all around the world unite once every 4 years to participate in the highly celebrated athletic competition. It is a worldwide proving ground for athletes and a chance for them to represent their countries.

International Olympic Committee (IOC) stated that the exact date of the first game is unknown, but the date of 776 BC is often cited [1]. It was held every 4 years at Olympia and thus acquired the name Olympics Games. However, the rise in power of the Romans and their influence upon Ancient Greece eventually resulted at the end of the event in around 400AD. The first modern Olympic Games were held in Athens in 1896 after Pierre de Coubertin announced his desire to resurrect the Olympic Games in 1894 [2].

With only 14 participating countries in 1896, the Olympics has seen a huge growth ever since; in 2020, 206 countries participated the Summer Olympics [3]. The types of games have expanded tremendously as the event is divided into two seasonally themed halves: Summer Olympics for fair-weather sports and Winter Olympics for cold-weather sports.

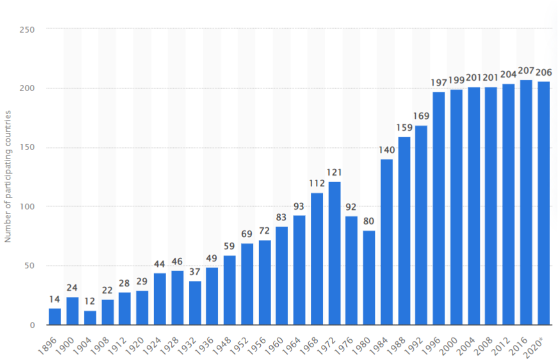


Figure 1: Number of participating countries in the Summer Olympics 1896-2021

However, the game wasn't always an inclusive arena. It wasn't until 1900 that women were allowed to compete and out of 997 athletes, 22 women competed in five sports: tennis, sailing, croquet, equestrianism, and golf [4]. Women competed for the first time in every event at the London 2012 Olympic Games and the Olympics course has changed ever since.

The Olympics has a powerful impact on the countries that participate, down to the host countries if done right. For example, the Sydney 2000 Olympic Games have been praised by experts as the best-organized Olympics in modern history, leaving behind a legacy of improved environmental conditions, a practical new transportation system, real estate growth, and an increase in tourism [5].

In addition to the impact on the countries themselves, the athletes are given huge incentives if they succeed to bring back gold medals. Hong Kong awarded 5 million Hong Kong dollars (about $642,000 at the current exchange rate) to Cheung Ka-long, a fencer who won gold in men's foil in Tokyo 2020 [6]. This makes both the countries and athletes all over the world compete ultimately and bring as many gold medals home as they can.

When the stakes are this high, it's crucial to observe what it takes for a nation to be the most successful at winning medals. One would think that since it was founded in Greece, it would have the most medals. It turns out that the United States won the most medals overall, followed by the United Kingdom, Germany, and France [7]. Without historical data, it would be impossible to track and study how the winning countries won their medals. Could it be that the way to win the Olympics is to first be a first-world country? Or is it because of the home advantages as seen in major professional team sports [8]?

This is where sports analytics comes in. The author was fortunate enough to have found an archive of the modern Olympic Games, recording the games from Athens 1896 to Rio 2016 [9]. The author would like to dive deeper into Olympics history and find the factors for winning the most medals.

The objective of this project is to:

- Present the descriptive statistics of the Olympics history.
- Determine if the GDP, climate, female to male ratio in a team of a country influences the medal tally.
- Determine if hosting the game influences the medal tally.
- Apply multiple linear regression to predict how many medals would the country win.

II. Literature Review

Not a lot of sports analytics study about the Olympics has been published. The author found that most of the published works are related to the economic side of the Olympics and not about evaluating how to tally the most medals in the games. In addition, most of them who are significant to the author's objective stopped after performing exploratory data analysis (EDA).

A well-done study was conducted by Pradhan et. al in 2021 where they performed EDA using the same dataset the author found [10]. Their main goal was to only analyze and visualize the various factors such as players' demographics. Pradhan et. al's work has helped the author to gain a sense of how to visualize the massive dataset so that the reader can understand the evolution of the Olympics better.

Another important study that helped the author to arrive at the hosting impact was done by Balmer et. al in 2010 [11]. They conducted research to assess home advantage in the Winter Olympics from 1908-1988 and applied linear regression to the dataset. They found that figure skating, freestyle skiing, ski jumping, alpine skiing, and short-track speed skating all displayed some signs of home advantage with $p = 0.037$. This greatly helped the author's suspicion of the hosting advantage as a factor in the medal tally.

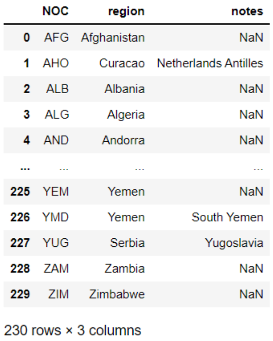
III. Data

The data set is available on Kaggle and was scraped by Griffin 4 years ago from a sports tracker website [9] [12]. The data set is available in .csv format which contains 271,116 rows and 15 columns. The list of the variables can be seen in Table I.

Variable	Details
ID	Unique Identifier
Name	Athlete's Name
Sex	M or F
Age	Integer
Height	In centimeters
Weight	In kilograms
Team	Team's Name
NOC	National Olympic Committee (3 letter code)
Games	Year and Season
Year	Integer
Season	Summer or Winter
City	Host City
Sport	Sport
Event	Event
Medal	Gold, Silver, Bronze, or NA

Table I: Variables in the Data Frame

In addition to the main data, there is a supporting data for the NOC variable. NOC is a label for where the athlete is from. For example, the NOC for China is CHN and Denmark is DEN. This will be useful later on since it's easier to match variables with short codes such as NOC's, rather than using the actual country names.

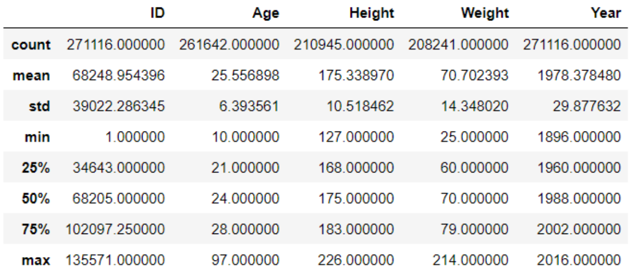


	NOC	region	notes
0	AFG	Afghanistan	NaN
1	AHO	Curacao	Netherlands Antilles
2	ALB	Albania	NaN
3	ALG	Algeria	NaN
4	AND	Andorra	NaN
...	...	...	...
225	YEM	Yemen	NaN
226	YMD	Yemen	South Yemen
227	YUG	Serbia	Yugoslavia
228	ZAM	Zambia	NaN
229	ZIM	Zimbabwe	NaN

230 rows × 3 columns

Figure 2: The NOC Dataframe

Even though we have 271,116 rows, there are only 135,571 unique athletes since an athlete can compete in a different branch of sports. The summary statistics of the numeric variables can be seen in Figure 3. The clean data frame can be seen in the notebook file.



	ID	Age	Height	Weight	Year
count	271116.000000	261642.000000	210945.000000	208241.000000	271116.000000
mean	68248.954396	25.556898	175.338970	70.702393	1978.378480
std	39022.286345	6.393561	10.518462	14.348020	29.877632
min	1.000000	10.000000	127.000000	25.000000	1896.000000
25%	34643.000000	21.000000	168.000000	60.000000	1960.000000
50%	68205.000000	24.000000	175.000000	70.000000	1988.000000
75%	102097.250000	28.000000	183.000000	79.000000	2002.000000
max	135571.000000	97.000000	226.000000	214.000000	2016.000000

Figure 3: Summary Statistics of Numerical Variables

The author needs to have another set of data since one of the objectives is to observe the relationship between GDP and medal tally. Fortunately, the author found another data set from Kaggle that was created by Chadha in 2018 [13]. This data set contains 222 countries' GDP values in USD, from 1960-2020. Since we don't have the GDP values from the year before that, we're just going to look at the medal tally starting from 1960.

IV. Methodology

A. *Pearson's Correlation Coefficient*

In this project, we would like to see if there's any relationship between GDP, climate, the female-to-male ratio in a team of a country, and hosting the game toward medal tally. To check this, we need to find their correlation.

Pearson's correlation coefficient is a way to evaluate sample data as evidence that a linear relationship exists between two quantitative variables [14]. This statistic is used to reject the null hypothesis: there is no association between two random variables X and Y.

The $X's$ are then GDP, climate, and female to male ratio while the Y is the number of medals collected from Olympics 1960-2016.

Define the correlation as $\rho_{X,Y}$. Then,

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \qquad (1)$$

where $Cov(X,Y)$ is the covariance between X and Y and σ_i is the standard deviation for $i = X,Y$. If the correlation is 1 (-1), then it means there is a perfect positive (negative) relationship; if 1 unit of X goes up, 1 unit of Y goes up (down), and vice versa. If it's 0, then there is no relationship between them.

Of course, we don't need to calculate this manually since this formula is available in pandas library:

```
# To show the correlation matrix of a data frame (numerical)
df.corr()
```

Furthermore, we need to adjust the climate and hosting variable since it is not quantitative by nature. The author will use a binary variable instead. If the country does not have a winter season, assign $clim = 0$, otherwise $clim = 1$. Since every country has summer/fair-weather climates, but not all have winter. Hypothetically, a country that has a winter season should do better in Winter Olympics than a country that doesn't. Same goes for hosting variable; $host = 0$ for not hosting and $host = 1$ for hosting.

B. *Multiple Linear Regression*

Linear regression with more than 1 predictor is called a multiple linear regression. For this project, the author set the response variable to be the number of medals collected and the predictors are GDP, climate, female-to-male ratio in the team, and whether or not they're hosting the Olympics. Thus, the model is

$$Y = \beta_0 + \beta_1 \times GDP + \beta_2 \times clim + \beta_3 \times ratio + \beta_4 \times host \qquad (2)$$

Using the least squares method, we would be able to find the β_i's that minimizes $RSS = \sqrt{(\hat{Y} - Y)^2}$, with the solution:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \qquad (3)$$

where X is the predictors' matrix and Y is the vector with the response variable as its entries [15].

REFERENCES

[1] I. O. Comittee, “What is the origin of the olympics games?” 2021, last accessed on 23 October 2022. [Online]. Available: https://olympics.com/ioc/faq/history-and-origin-of-the-games/what-is-the-origin-of-the-olympic-games

[2] ——, “Who was pierre de coubertin?” 2021, last accessed on 23 October 2022. [Online]. Available: https://olympics.com/ioc/faq/history-and-origin-of-the-games/who-was-pierre-de-coubertin

[3] S. R. Department, “Number of participating countries in the summer olympics 1896-2021,” 2022, last accessed on 24 October 2022. [Online]. Available: https://www.statista.com/statistics/280462/summer-olympics-1896-2012-number-of-participating-countries/

[4] I. O. Comittee, “When did women first compete in the olympic games?” 2021, last accessed on 24 October 2022. [Online]. Available: https://olympics.com/ioc/faq/history-and-origin-of-the-games/when-did-women-first-compete-in-the-olympic-games

[5] J. Mossop, “Sydney has set the highest standards for future hosts,” 2000, last accessed on 24 October 2022. [Online]. Available: https://www.telegraph.co.uk/sport/4772754/Sydney-has-set-the-highest-standards-for-future-hosts.html

[6] B. Knight, “These 12 countries will pay their olympians six-figure bonuses for winning gold medals,” 2022, last accessed on 24 October 2022. [Online]. Available: https://www.forbes.com/sites/brettknight/2022/02/04/these-12-countries-will-pay-their-olympians-six-figure-bonuses-for-winning-gold-medals/

[7] I. O. Comittee, “Records and medals,” 2021, last accessed on 24 October 2022. [Online]. Available: https://olympics.com/ioc/documents/olympic-games/records-and-medals

[8] R. Pollard, “Home advantage in soccer: A retrospective analysis,” *Journal of sports sciences*, vol. 4, pp. 237–48, 02 1986.

[9] R. Griffin, “120 years of olympic history: athletes and results,” 2018, last accessed on 24 October 2022. [Online]. Available: https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results

[10] R. Pradhan, K. Agrawal, and A. Nag, “Analyzing evolution of the olympics by exploratory data analysis using r,” *IOP Conference Series: Materials Science and Engineering*, vol. 1099, no. 1, p. 012058, Mar. 2021. [Online]. Available: https://doi.org/10.1088/1757-899x/1099/1/012058

[11] N. J. Balmer, A. M. Nevill, and A. M. Williams, “Home advantage in the winter olympics (1908-1998),” *Journal of Sports Sciences*, vol. 19, no. 2, pp. 129–139, Jan. 2001. [Online]. Available: https://doi.org/10.1080/026404101300036334

[12] T. S. Reference, “Sports stats, fast, easy, and up-to-date,” 2018, last accessed on 22 October 2022. [Online]. Available: https://www.sports-reference.com/

[13] S. Chadha, “Country wise gdp data,” 2018, last accessed on 25 October 2022. [Online]. Available: https://www.kaggle.com/datasets/chadalee/country-wise-gdp-data

[14] D. LeBlanc, *Statistics: Concepts and Applications for Science*. Jones and Bartlett, 2004. [Online]. Available: https://books.google.ca/books?id=gtawVU0oZFMC

[15] D. W. T. H. James, Gareth and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*. Springer, 2021.