# Olympic Medal Analysis - Exploratory Data Analysis

Muhammad Y. Khan, Ge Zhang, Chao Wang, Saud A. Abbasi

## Introduction

This Exploratory Data Analysis (EDA) investigates the Olympic medal dataset and offers a thorough examination of Olympic history from Athens in 1896 to Rio in 2016 using data gathered from a variety of sources. The main source of the data is Heesoo Park's (Park 2018) Kaggle dataset, "120 years of Olympic history: athletes and results." This large dataset, which covers both summer and winter games, offers a comprehensive historical record of Olympic participants and their performances. In this EDA a dataset that was selected by the TidyTuesday project (R. C. 2024. T. Dataset 2024) is utilized, which provides a subset of this data that may be used to practice data visualization and analysis. The same Kaggle source is used to create this TidyTuesday dataset.

The goal of the R4DS Online Learning Community's TidyTuesday project is to give data enthusiasts access to real-world datasets so they may hone their data wrangling and visualization capabilities (Wickham and Contributors 2023).By revisiting this Olympic dataset (previously featured in 2021 (R. C. 2021. T. Dataset 2021)), the project allows for continued exploration of athletic achievements.

## Data Sources

The primary data is sourced from the Kaggle dataset by Heesoo Park:

https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results/data

A supplementary dataset is obtained from the TidyTuesday GitHub repository:

https://raw.githubusercontent.com/rfordatascience/tidytuesday/main/data/2024/2024-08-06/olympics.csv

Reference to the analysis performed by Heesoo Park:

https://www.kaggle.com/code/heesoo37/olympic-history-data-a-thorough-analysis

## Data Description

A vast amount of data regarding Olympic athletes and their performances from Athens 1896 to Rio 2016 is included in the merged databases. Below is a thorough explanation of each column:

- **id**: Unique identifier for each athlete.
- **name**: Athlete's full name.
- **sex**: Athlete's gender (M/F).
- **age**: Athlete's age at the time of the Olympic event.
- **height**: Athlete's height in centimeters.
- **weight**: Athlete's weight in kilograms.
- **team**: The team or nation the athlete represented.
- **noc**: National Olympic Committee code.
- **games**: Specific Olympic Games (e.g., "2016 Summer").
- **year**: Year of the Olympic Games.
- **season**: Season of the Games (Summer or Winter).
- **city**: Host city of the Olympic Games.
- **sport**: Sport in which the athlete competed.
- **event**: Specific event within the sport.
- **medal**: Type of medal won (Gold, Silver, Bronze, or NA if no medal was won).

Note that both the Summer and Winter Olympics are included in the dataset. The Summer and Winter Games were held in the same year until 1992. Following 1992, they were spaced out, with the Winter Games taking place in 1994, followed by the Summer Games in 1996, and so on. More detailed data on athlete participation and performance across the whole time span can be found in the Kaggle dataset.

### Unique Values in Categorical Variables

Let's look at how the TidyTuesday dataset's important category variables are distributed:

```r
#| label: unique-values

cat("Medals:", paste(unique(olympics$medal), collapse = ", "), "\n")
```

```
Medals: NA, Gold, Bronze, Silver
```

```r
cat("Seasons:", paste(unique(olympics$season), collapse = ", "), "\n")
```

```
Seasons: Summer, Winter
```

```
cat("Sexes:", paste(unique(olympics$sex), collapse = ", "), "\n")
```

```
Sexes: M, F
```

These findings help us understand the gender categories utilized in the dataset, the Olympic seasons, and the medal categories.

### Potential Research Questions

Given the richness of this dataset, and inspired by the questions explored by Heesoo Park, numerous intriguing research questions emerge:

1. Does the age distribution of Olympic medalists differ by gender and sport, and how has it changed over time?
2. Does an athlete's success in a particular sport correlate with their physical characteristics (weight, height), and if so, how has this relationship changed over time?
3. How has the proportion of women in the Olympics changed since 1896, and how do medal distributions vary by gender in various sports and eras?
4. In light of political and economic pressures, which nations have seen the biggest increases in medal counts throughout time, and what reasons might account for these gains?
5. After adjusting for variables like population and economic development, are there any discernible patterns in the host nations' performance, and is there proof of a "host country advantage"?
6. In light of geographic considerations, which sports are more popular in each season and how do Winter Olympic participants differ from Summer Olympic athletes?
7. Do different countries' medal counts and team sizes have a link? Does this relationship alter depending on the sport or region? Does growing a team size have diminishing returns?
8. How do summer and winter games differ in terms of performance (medals won)?

## Bibliography

Dataset, R4DS Community: 2021 TidyTuesday. 2021. "TidyTuesday: Olympic Dataset (2021)." https://github.com/rfordatascience/tidytuesday/blob/main/data/2021/2021-07-27/readme.md.

Dataset, R4DS Community: 2024 TidyTuesday. 2024. "TidyTuesday: Olympic Dataset (2024)." https://github.com/rfordatascience/tidytuesday/blob/main/data/2024/2024-08-06/readme.md.

Park, Heesoo. 2018. “120 Years of Olympic History: Athletes and Results Dataset.” https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results/data.

Wickham, Hadley, and Contributors. 2023. “R for Data Science: Quarto Edition.” https://r4ds.hadley.nz/quarto.html.