

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 2018a) 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/tidyuesday/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:

```
#| label: unique-values

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

```
Medals:  NA, Gold, Bronze, Silver
Seasons:  Summer, Winter
Sexes:   M, F
```

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

Initial Observations

Based on Heesoo Park's analysis (Park 2018b) and our preliminary investigation of the TidyTuesday dataset, the following can be concluded:

1. **Data Range:** The dataset covers 120 years of Olympic history, from 1896 to 2016. Park points out that trends in which nations and sports predominate can be seen by looking at participation and medal tallies throughout time.

2. **Demographics of Athletes:**

From extremely young athletes to people in their 60s or even 70s, the age range is rather broad. Park's analysis recommends looking into the age distributions of various genders and sports.

- The wide range of sports represented probably contributes to the significant diversity in height and weight. Additional sport-specific analysis may uncover trends pertaining to the physical prerequisites for success.

3. **Gender Representation:** The dataset contains both male and female athletes. Park draws attention to the growing number of female athletes over time.

4. **Geographic Diversity:** The 'team' and 'noc' columns show that involvement came from a wide range of nations. Park's illustration of the medal distribution per nation offers insightful background information.

5. **Diversity at the Olympics:**

- There are representatives from both the summer and winter Olympics.
- There are several host cities. Potential "host country advantage," as investigated by Park, could be discovered by comparing the performance of host nations to their average performance.

6. **Sports and activities:** The Olympics' diversity is demonstrated by the large number of sports and specialized activities. Finding the sports with the most medals and participation would be a good place to start.

7. **Medal Information:** Silver, Bronze, and Gold medals are noted, although several entries indicate NA, which means non-medalists are also listed. A thorough examination of Olympic participation requires an understanding of the percentage of non-medalists. Medal ratios, which are included in Park's research, offer additional information.

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. 2018a. “120 Years of Olympic History: Athletes and Results Dataset.” <https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results/data>.
- . 2018b. “Thorough Analysis of Olympic History Data.” <https://www.kaggle.com/code/heesoo37/olympic-history-data-a-thorough-analysis>.
- Wickham, Hadley, and Contributors. 2023. “R for Data Science: Quarto Edition.” <https://r4ds.hadley.nz/quarto.html>.