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/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.

Initial Data Exploration

Firstly the data is loaded into R. For the sake of conciseness, its important to concentrate the preliminary investigation on the TidyTuesday dataset. It is crucial to keep in mind that both datasets' ((R. C. 2021. T. Dataset 2021) and (R. C. 2024. T. Dataset 2024)) source data covers the years 1896–2016 from Athens.

```r
library(tidyverse)
library(janitor)
```

```r
olympics <- read_csv(paste0(
  "https://raw.githubusercontent.com/rfordatascience/",
  "tidytuesday/main/data/2024/2024-08-06/olympics.csv"
))
olympics <- clean_names(olympics)
```

Glimpse of the Data

Let’s take a look at the first few rows and the structure of the TidyTuesday dataset.

```
glimpse(olympics)
```

```
Rows: 271,116
Columns: 15
$ id     <dbl> 1, 2, 3, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, ~
$ name   <chr> "A Dijiang", "A Lamusi", "Gunnar Nielsen Aaby", "Edgar Lindenau~
$ sex    <chr> "M", "M", "M", "M", "F", "F", "F", "F", "F", "F", "M", "M", "M"~
$ age    <dbl> 24, 23, 24, 34, 21, 21, 25, 25, 27, 27, 31, 31, 31, 31, 33, 33,~
$ height <dbl> 180, 170, NA, NA, 185, 185, 185, 185, 185, 185, 188, 188, 188, ~
$ weight <dbl> 80, 60, NA, NA, 82, 82, 82, 82, 82, 82, 75, 75, 75, 75, 75, 75,~
$ team   <chr> "China", "China", "Denmark", "Denmark/Sweden", "Netherlands", "~
$ noc    <chr> "CHN", "CHN", "DEN", "DEN", "NED", "NED", "NED", "NED", "NED", ~
$ games  <chr> "1992 Summer", "2012 Summer", "1920 Summer", "1900 Summer", "19~
$ year   <dbl> 1992, 2012, 1920, 1900, 1988, 1988, 1992, 1992, 1994, 1994, 199~
$ season <chr> "Summer", "Summer", "Summer", "Summer", "Winter", "Winter", "Wi~
$ city   <chr> "Barcelona", "London", "Antwerpen", "Paris", "Calgary", "Calgar~
$ sport  <chr> "Basketball", "Judo", "Football", "Tug-Of-War", "Speed Skating"~
$ event  <chr> "Basketball Men's Basketball", "Judo Men's Extra-Lightweight", ~
$ medal  <chr> NA, NA, NA, "Gold", NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
```

Summary Statistics

Now, let’s generate summary statistics for the numerical columns in the TidyTuesday dataset.

```
summary(olympics)
```

```
       id             name               sex                 age       
 Min.   :     1   Length:271116      Length:271116      Min.   :10.00  
 1st Qu.: 34643   Class :character   Class :character   1st Qu.:21.00  
 Median : 68205   Mode  :character   Mode  :character   Median :24.00  
 Mean   : 68249                                         Mean   :25.56  
```

```
3rd Qu.:102097                                         3rd Qu.:28.00
Max.   :135571                                         Max.   :97.00
                                                       NA's   :9474
    height          weight          team               noc
Min.   :127.0   Min.   : 25.0   Length:271116      Length:271116
1st Qu.:168.0   1st Qu.: 60.0   Class :character   Class :character
Median :175.0   Median : 70.0   Mode  :character   Mode  :character
Mean   :175.3   Mean   : 70.7
3rd Qu.:183.0   3rd Qu.: 79.0
Max.   :226.0   Max.   :214.0
NA's   :60171   NA's   :62875
   games                 year          season              city
Length:271116      Min.   :1896   Length:271116      Length:271116
Class :character   1st Qu.:1960   Class :character   Class :character
Mode  :character   Median :1988   Mode  :character   Mode  :character
                   Mean   :1978
                   3rd Qu.:2002
                   Max.   :2016

   sport              event              medal
Length:271116      Length:271116      Length:271116
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character
```

From the summary statistics the following is observed:

Athlete Demographics

- The dataset includes over **135,000 unique athletes**, based on the `id` column.
- **Age Distribution**:
  - The youngest recorded athlete was **10 years old**, while the oldest was **97**.
  - The **median age is 24**, and the **mean is 25.56**, indicating a slight **right skew** due to older athletes.
  - Most athletes are in their **20s**, with an interquartile range (IQR) of **21 to 28**.
  - **9,474 missing age values**, which may need to be addressed.

Physical Attributes

- **Height** ranges from **127 cm to 226 cm**, with a **median of 175 cm**.
- **Weight** varies from **25 kg to 214 kg**, with a **median of 70 kg**.
- **Significant missing values**:
  - **60,171 missing height values**
  - **62,875 missing weight values**
- The missing data might **impact analyses** involving athlete body metrics.

Olympic Games Timeline

- The dataset covers **Olympics from 1896 to 2016**.
- The **median year is 1988**, meaning half of the events occurred before this year and half after.
- The **mean year (1978)** being close to the median suggests **a balanced distribution of data** over time.

Team Representation

- The dataset contains **various teams and National Olympic Committees (NOCs)**.
- Since these are categorical variables, further analysis can help understand **country participation trends**.

Medal Data

- The `medal` column indicates whether an athlete won a **Gold, Silver, or Bronze medal**.
- Many entries contain **NA values**, suggesting that a **large number of athletes participated but did not win a medal**.
- Exploring medal-winning trends over time could provide insights into **which countries or athletes have dominated the Olympics**.

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.

Unique Values in Categorical Variables

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

```r
olympics |> distinct(medal)
```

```
# A tibble: 4 x 1
  medal
  <chr>
1 <NA>
2 Gold
```

```
3 Bronze
4 Silver
```

```r
olympics |> distinct(season)
```

```
# A tibble: 2 x 1
  season
  <chr>
1 Summer
2 Winter
```

```r
olympics |> distinct(sex)
```

```
# A tibble: 2 x 1
  sex
  <chr>
1 M
2 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. 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.