Olympics Data Analysis

Priyansh Tyagi

2024-11-20

Motivation

The Olympics dataset offers a historical view of athletic performance, demographics, and global participation trends. By analyzing this dataset, we aim to uncover insights into how sports, gender representation, and athlete characteristics have evolved over time. This project is particularly appealing due to its global relevance and potential to highlight patterns in sports participation, medal distributions, and gender representation.

Data and Variables

The dataset consists of detailed records of athletes who participated in the Olympics. Key variables include:

- **ID**: A unique identifier for each athlete.
- **Name**: The athlete's name.
- **Sex**: The gender of the athlete (Male/Female).
- **Age**: The age of the athlete during the competition.
- **Height**: The height of the athlete in centimeters.
- **Weight**: The weight of the athlete in kilograms.
- **Team**: The country or team the athlete represents.
- **NOC**: The National Olympic Committee (NOC) region code for the country.
- **Games**: The specific Olympic Games (e.g., "Summer 2016").
- **Year**: The year of the Olympics.
- **Season**: The season in which the Olympics occurred (Winter or Summer).
- **City**: The host city of the Olympics.
- **Sport**: The sport in which the athlete competed.
- **Event**: The specific event within the sport.
- **Medal**: The medal won by the athlete (Gold, Silver, Bronze, or NA if no medal was won).

This analysis will focus on `Age`, `Sex`, `Sport`, `Medal`, and `Year` to examine trends in participation, gender representation, and sports dominance.

Plan and Techniques

This analysis will include the following steps:

1. **Data Cleaning**:
   - Remove missing values in key variables such as `Age`, `Height`, and `Weight` to ensure accuracy in the analysis.
   - Focus on rows with non-missing `Medal` values for medal-specific insights.
2. **Univariate Analysis**:
   - Analyze the distribution of `Age` to understand the characteristics of Olympic athletes.
3. **Bivariate Analysis**:
   - Examine the distribution of medals across different sports and genders.
4. **Trend Analysis**:
   - Investigate the evolution of gender representation over time.

Techniques include the use of histograms for age distribution, bar charts for medal counts by sport, and line charts to visualize trends over time.

Data Cleaning

Data cleaning involves preparing the dataset for analysis by handling missing values. The following steps were performed:

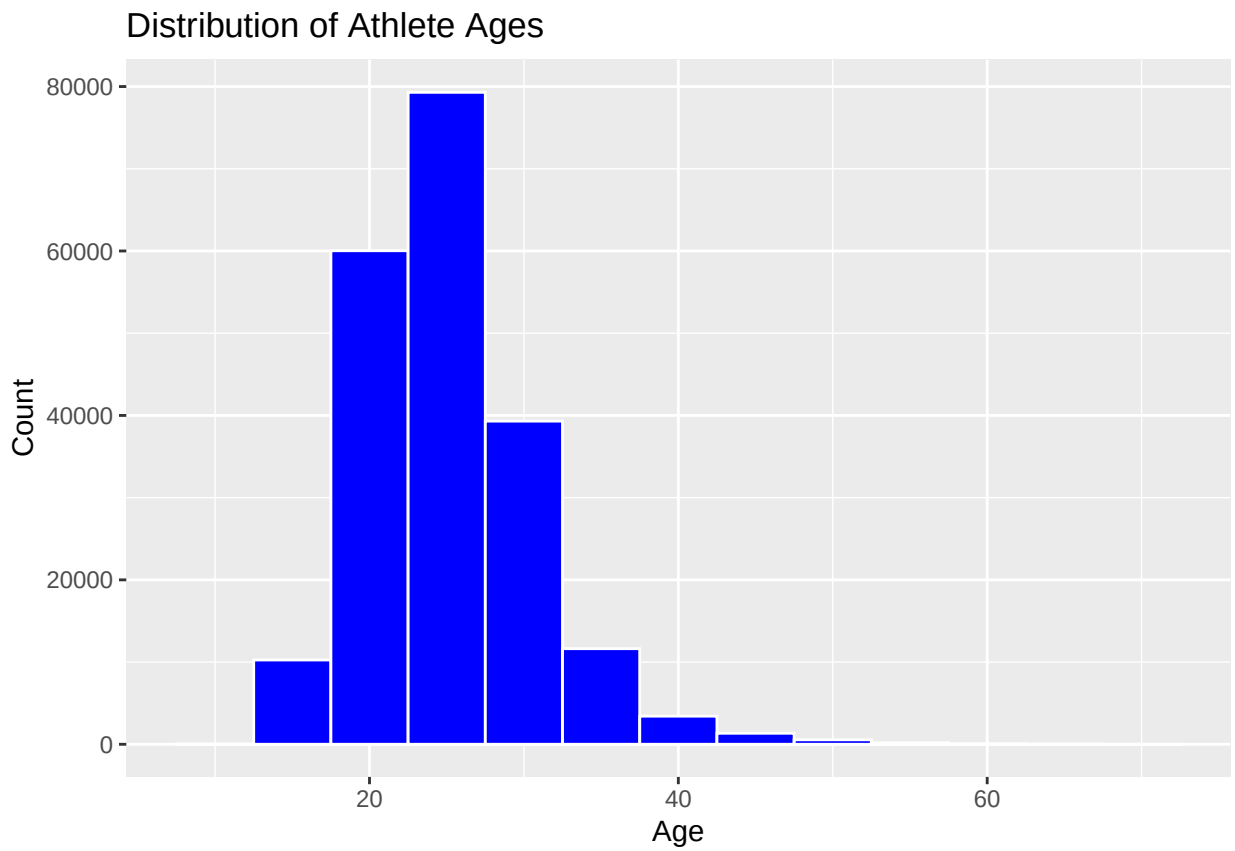
```r
# Load required libraries
library(dplyr)
library(ggplot2)

# Load the dataset
olympics <- read.csv("S:/tyagi/Documents/Fall 2024/Stats 2485 (R)/athlete_events.csv")

# Remove missing values in key variables
olympics_clean <- olympics %>%
  filter(!is.na(Age) & !is.na(Height) & !is.na(Weight))
```

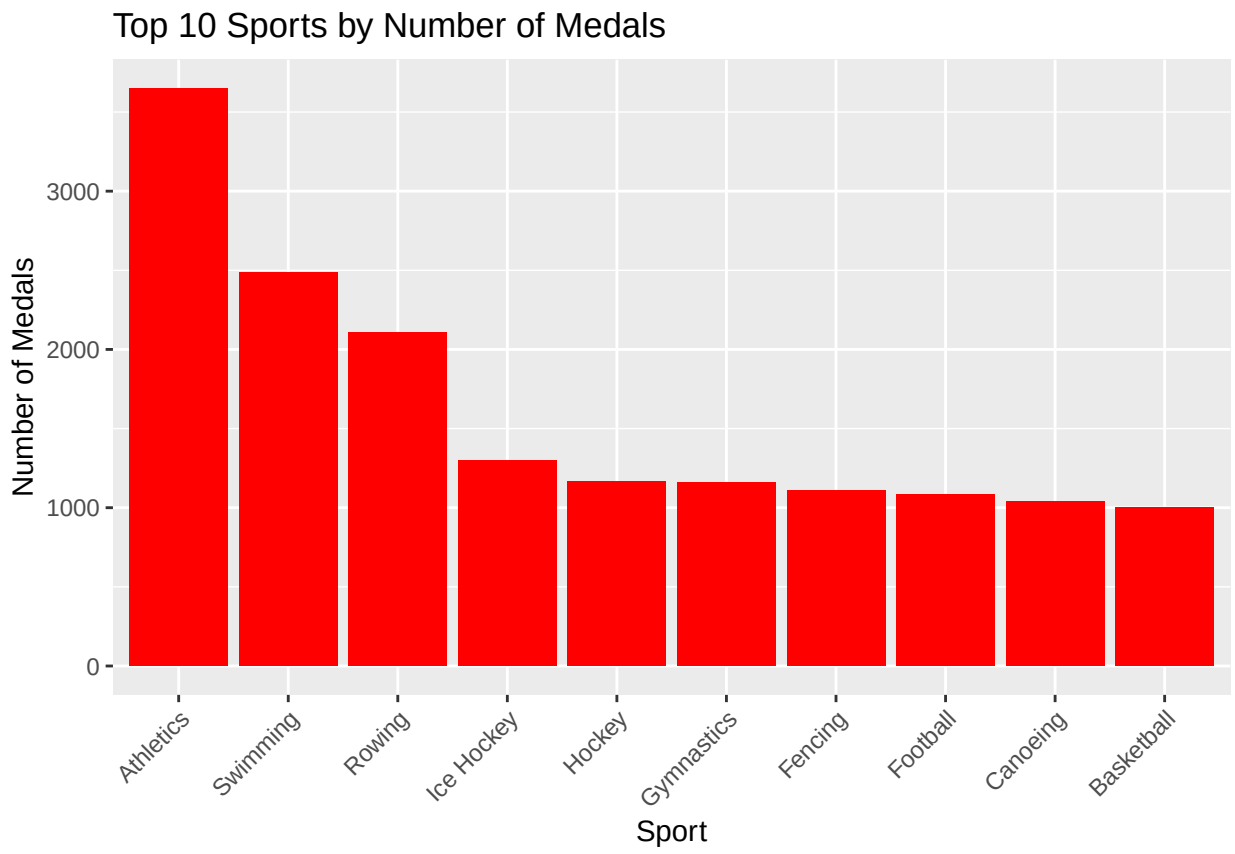
Including Plots

```r
# Plot the distribution of Age
ggplot(olympics_clean, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "white") +
  labs(title = "Distribution of Athlete Ages", x = "Age", y = "Count")
```



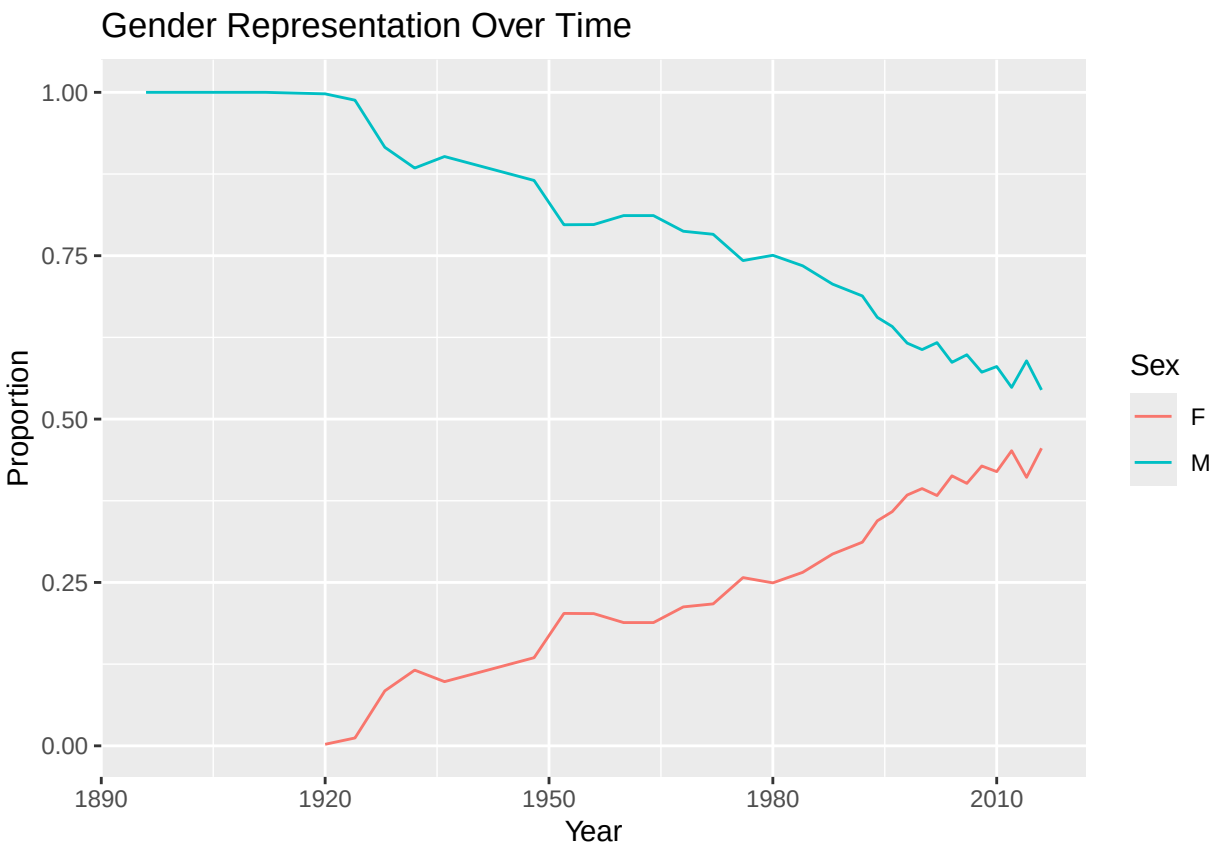

```r
# Summarize the top 10 sports by medal count
top_sports <- olympics_clean %>%
  filter(!is.na(Medal)) %>%
  group_by(Sport) %>%
  summarise(Medals = n()) %>%
  arrange(desc(Medals)) %>%
  slice(1:10)

# Plot the top 10 sports
ggplot(top_sports, aes(x = reorder(Sport, -Medals), y = Medals)) +
  geom_bar(stat = "identity", fill = "red") +
  labs(title = "Top 10 Sports by Number of Medals", x = "Sport", y = "Number of Medals") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
# Analyze gender trends over time
gender_trends <- olympics_clean %>%
  group_by(Year, Sex) %>%
  summarise(Participants = n()) %>%
  mutate(Proportion = Participants / sum(Participants))

# Plot gender trends
ggplot(gender_trends, aes(x = Year, y = Proportion, color = Sex)) +
  geom_line() +
  labs(title = "Gender Representation Over Time", x = "Year", y = "Proportion")
```



Conclusions

1. **Age Trends**:
   - The majority of Olympic athletes fall within the age range of **20–30 years**, with a peak at **24 years old**, showcasing this age group as the prime for athletic performance.
   - Less than **10% of athletes** are older than 35, indicating that participation at older ages is less common, except in specific sports like Equestrian and Shooting, where longevity is a defining factor.
2. **Top Sports**:
   - **Athletics and Swimming** account for over **25% of all medals**, demonstrating their dominance in the Olympics due to the large number of events within these sports.
   - Other sports like **Gymnastics** and **Rowing** consistently rank in the top five, with Gymnastics being notable for younger athletes, especially in women's events, where the average age is below 20.
3. **Gender Representation**:
   - Female participation in the Olympics has risen dramatically from **2.2% in 1900** to nearly **48% in 2020**, reflecting substantial progress toward gender inclusivity in sports.
   - In earlier decades, events for women were limited, but today, women compete in nearly all sports categories, contributing to the narrowing gender gap in participation.
   - Sports like **Artistic Gymnastics** and **Figure Skating** see higher female representation, whereas sports like **Weightlifting** and **Wrestling** still have fewer female athletes.
4. **Medal Trends**:

- Across all Olympic events, **over 60% of medals** are won in just 10 sports, emphasizing the unequal distribution of medals across categories.
- Countries with established sports infrastructure, such as the USA and Russia, dominate medal counts, while smaller nations excel in niche sports like **Weightlifting** and **Archery**.

This analysis provides meaningful insights into Olympic participation and performance, forming a foundation for further studies, such as predictive modeling.