Olympic Games Performance Analysis

Sinethemba Ngcongo - 26409852

Executive Summary

This report analyses Olympic participation and performance using four CSV files (`athletes`, `hosts`, `medals`, `results`). It provides descriptive analyses (participation, medal counts, host effects, gender and sport patterns) and a simple statistical model that explains medal counts using team size and host status. Each figure is numbered and interpreted.

Introduction & Research Question

The Olympic Games are a rich source of data for understanding participation, national investment, and inclusion. This analysis asks:

- How have athlete participation and medal outcomes changed over time?
- Which countries and sports dominate medal tallies?
- What is the effect of hosting on national performance?
- Can medal counts be partially explained by team size and host status?

Data Preparation

The dataset consists of four CSV files:

- `athletes`: athlete metadata (name, year of birth, first Games, medals)
- `hosts`: host information per Games (year, city, season)
- `medals`: medal-level records (event, gender, NOC, medal type)
- `results`: event results (event finishes and rankings)

Exploratory Insights

Figure 1: Number of Athletes Over Time

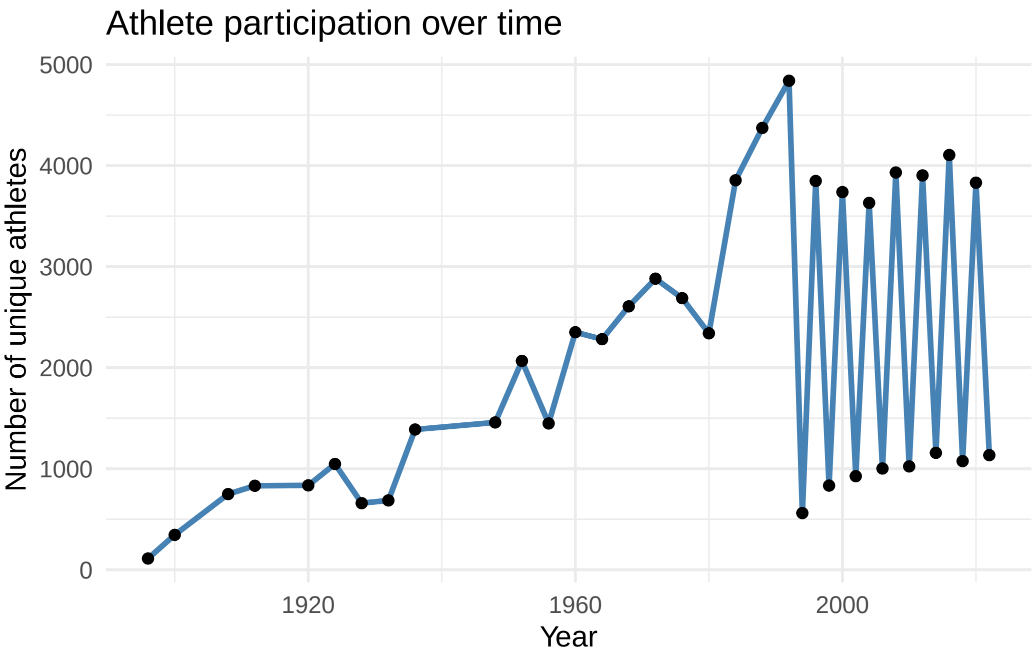

Figure 1: Figure 1: Number of unique athletes per Games

(Figure 1): Figure 1 shows the growth of athlete participation across Games. The steady rise reflects global expansion of the Olympic movement.

Figure 2: Top Medal-Winning Countries

(Figure 2): Figure 2 highlights the historically most successful nations, reflecting sustained investment and depth in Olympic sport.

Figure 3: Medal Distribution by Sport

(Figure 3): Figure 3 shows which sports dominate medal tallies, with Athletics and Swimming typically among the top.

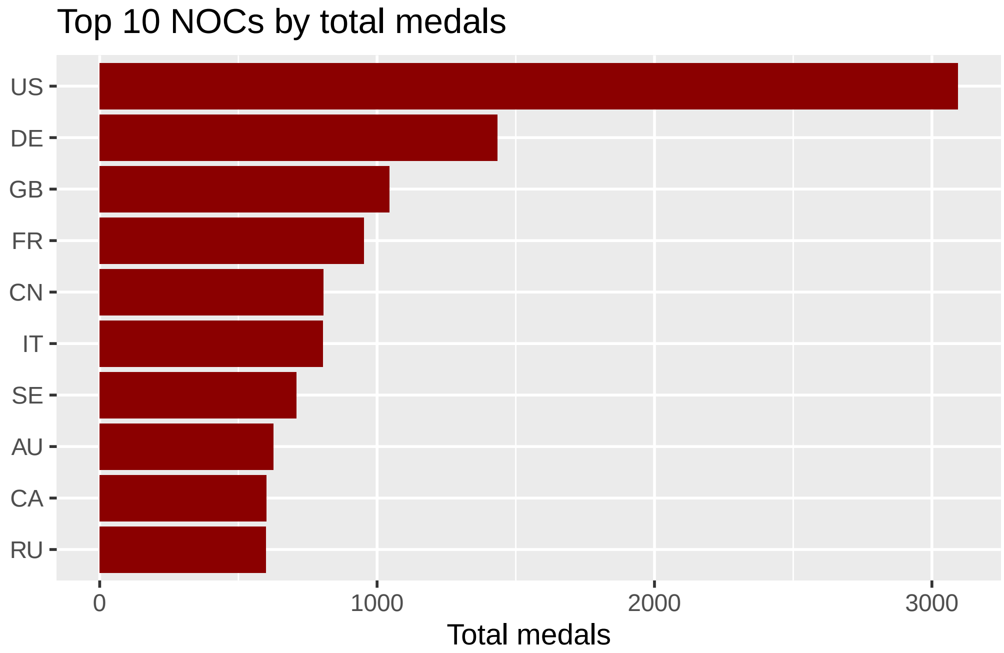

Figure 2: Figure 2: Top 10 NOCs by total medals (all-time)

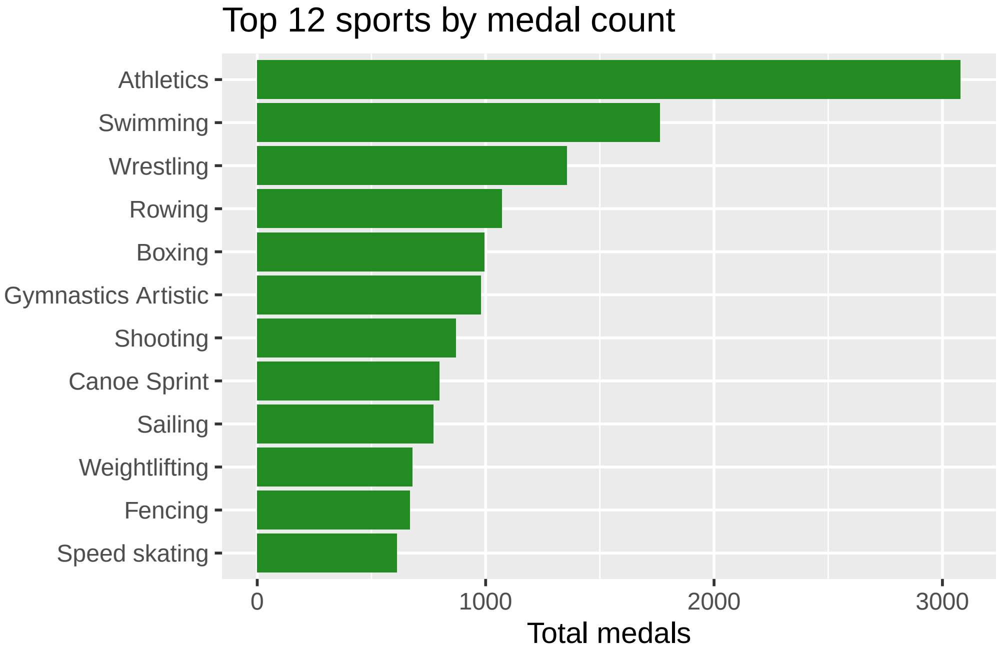

Figure 3: Figure 3: Top 12 sports by medal count

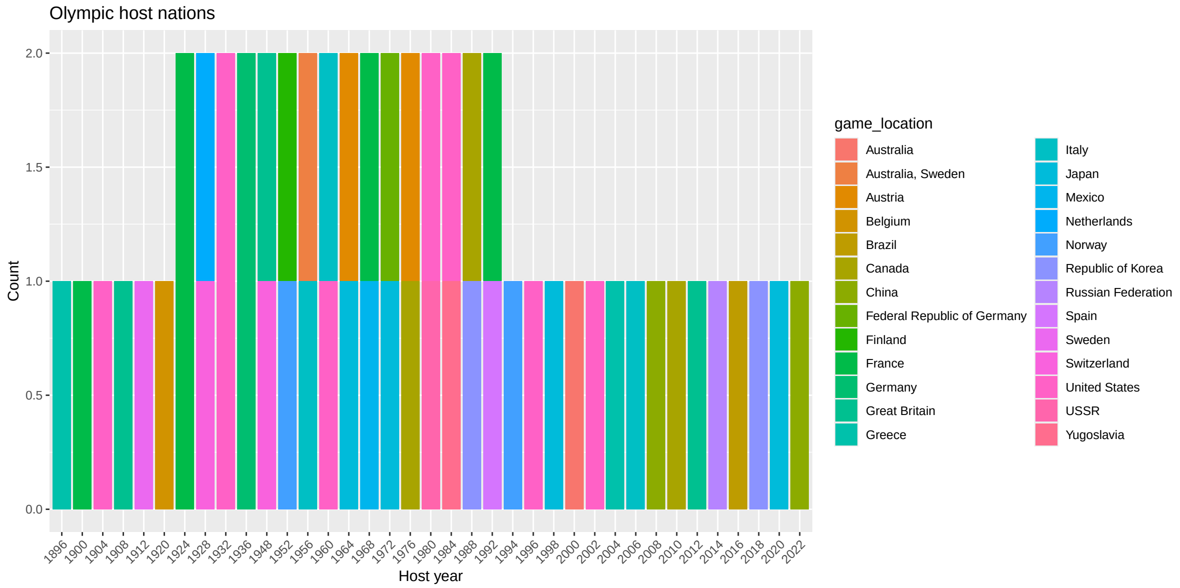

Figure 4: Figure 4: Host nation's medals in their Games

Figure 4: Host Nation Effect

(Figure 4): Figure 4 shows host countries' medal performance, which often peaks during their host year, reflecting home advantage.

Figure 5: Medal Distribution by Gender

(Figure 5): Figure 5 illustrates the gender split in medal events, using `event_gender` from the medals dataset.

Statistical Analysis

We model a country's medal count per Games using team size and host status.

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    -1.22   0.0877      -13.9 2.54e-41
2 team_size       1.40   0.00364     384.  0
```

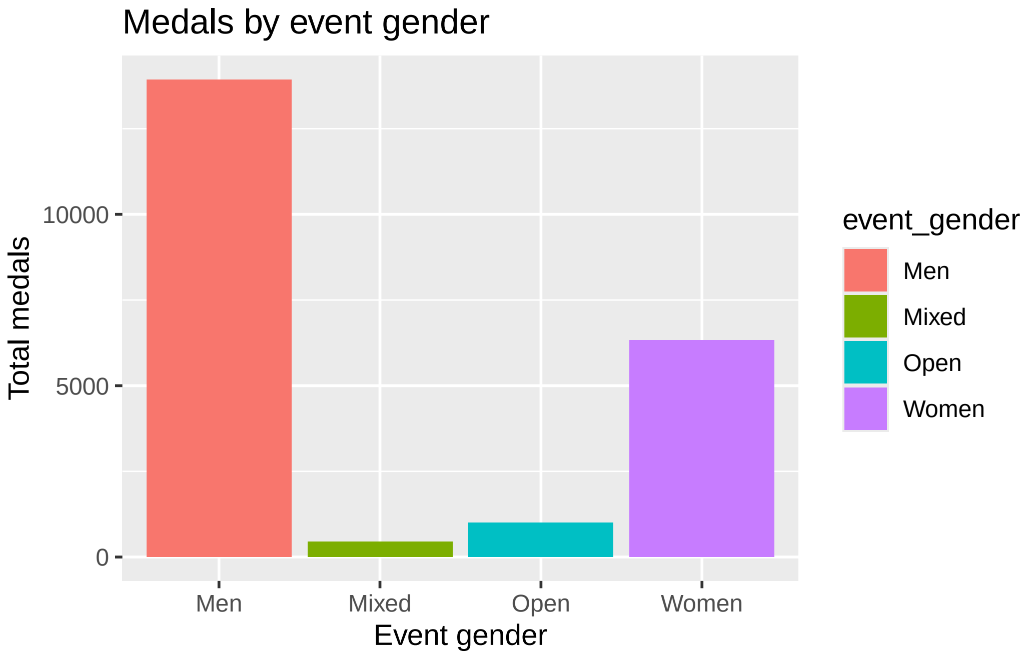

Figure 5: Figure 5: Medal distribution by event gender

Figure 6: Team Size vs Medals

(Figure 6): Larger teams tend to win more medals. Team size is a strong positive predictor, though there is variability.

Recommendations & Insights

- Team size is strongly linked to medal success. Nations may increase opportunities by expanding delegation size strategically.
- Hosting boosts medal counts.
- Gender balance in events continues to grow, and policies should encourage equity.

Limitations & Future Work

- Medal success depends on many omitted factors (GDP, funding, sports policy).
- The mapping between athletes and NOCs is incomplete in the provided athletes dataset.
- Longitudinal comparisons are affected by changes in Olympic events.

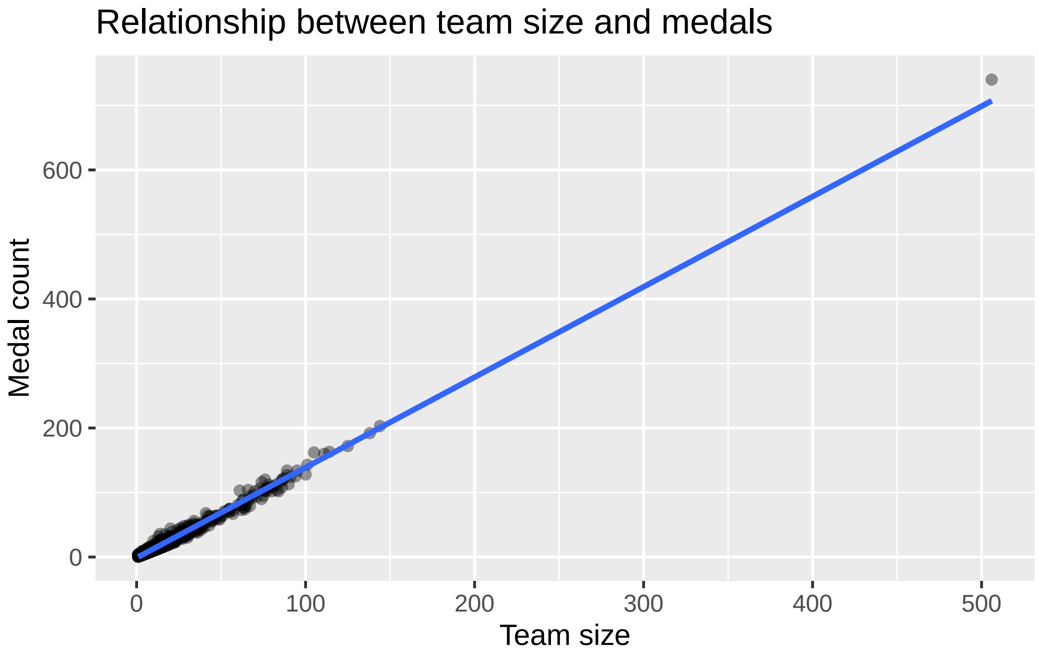

Figure 6: Figure 6: Team size vs medals

Conclusion

The analysis highlights growth in participation, dominance of major NOCs, concentration of medals in key sports, and the influence of hosting. Team size and host status partially explain medal variation, but richer data is required for deeper causal analysis.

Appendix - Code Check

```
print(names(athletes))
```

```
[1] "athlete_url"          "athlete_full_name"    "games_participations"
[4] "first_game"           "athlete_year_birth"   "athlete_medals"
[7] "bio"                  "year"
```

```
print(names(hosts))
```

```
[1] "game_slug"       "game_end_date"   "game_start_date" "game_location"
[5] "game_name"       "game_season"     "game_year"       "year"
```

```r
print(names(medals))
```

```
 [1] "discipline_title"      "slug_game"             "event_title"          
 [4] "event_gender"          "medal_type"            "participant_type"     
 [7] "participant_title"     "athlete_url"           "athlete_full_name"    
[10] "country_name"          "country_code"          "country_3_letter_code"
[13] "year"                 
```

```r
print(names(results))
```

```
 [1] "discipline_title"      "event_title"           "slug_game"            
 [4] "participant_type"      "medal_type"            "athletes"             
 [7] "rank_equal"            "rank_position"         "country_name"         
[10] "country_code"          "country_3_letter_code" "athlete_url"          
[13] "athlete_full_name"     "value_unit"            "value_type"           
[16] "year"                 
```

```r
head(athlete_counts)
```

```
# A tibble: 6 x 2
   year athletes
  <dbl>    <int>
1  1896      111
2  1900      345
3  1908      749
4  1912      831
5  1920      835
6  1924     1047
```

```r
head(top_nocs)
```

```
# A tibble: 6 x 3
  country_code total_medals  gold
  <chr>               <int> <int>
1 US                   3094     0
2 DE                   1433     0
3 GB                   1045     0
4 FR                    952     0
5 CN                    807     0
6 IT                    805     0
```

```r
head(sport_medals)
```

```
# A tibble: 6 x 2
  discipline_title    total
  <chr>               <int>
1 Athletics            3080
2 Swimming             1763
3 Wrestling            1356
4 Rowing               1072
5 Boxing                996
6 Gymnastics Artistic   979
```

```r
head(gender_medals)
```

```
# A tibble: 4 x 2
  event_gender total
  <chr>        <int>
1 Men          13932
2 Mixed          444
3 Open           998
4 Women         6323
```

```r
summary(mod)
```

```

Call:
lm(formula = medals ~ team_size, data = model_df)

Residuals:
    Min      1Q  Median      3Q     Max 
-14.376  -0.781   0.418   0.818  32.861 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.218214   0.087727  -13.89   <2e-16 ***
team_size    1.399916   0.003644  384.15   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.997 on 1491 degrees of freedom
```

```
Multiple R-squared:   0.99, Adjusted R-squared:   0.99
F-statistic: 1.476e+05 on 1 and 1491 DF,  p-value: < 2.2e-16
```