Olympic Performance and GDP Analysis

Arul Santoshi, Kyle Spaulding, Krish Chavan

Introduction

This report explores the relationship between national economic output (GDP) and Olympic performance in the Summer Olympics. Using cleaned and merged datasets, we analyze medal counts, efficiency metrics, and regression relationships to better understand how economic resources relate to athletic success.

Analytical Framework

Paradigms and Perspectives

This analysis adopts both **exploratory** and **confirmatory** data analysis approaches. Exploratory data analysis (EDA) allows us to discover patterns, detect outliers, and visualize trends without preconceived hypotheses. In contrast, confirmatory analysis tests specific hypotheses using regression models to quantify the GDP–medal relationship [@tukey1977exploratory].

We chose this dual approach because Olympic performance is a complex phenomenon requiring both discovery (e.g., identifying efficiency paradoxes) and hypothesis testing (e.g., quantifying GDP's predictive power). For coding, we adopt the **tidyverse paradigm** [@wickham2019welcome], which emphasizes readable data pipelines using `%>%` operators, consistent naming conventions, and human-readable function names. This approach enhances reproducibility and collaboration by making our data transformations explicit and easy to follow.

Data Sources and FAIR/CARE Principles

Data Provenance

Our analysis integrates two primary datasets:

1. **Summer Olympics Medal Data**: Scraped from publicly available Olympic records spanning 1960–2020 [@olympics_data_2024]. This dataset includes country names, years, and medal counts (Gold, Silver, Bronze).
2. **GDP Data**: Retrieved from the World Bank Open Data portal [@worldbank2024], providing annual GDP figures in current US dollars for all nations from 1960 onward.

FAIR and CARE Assessment

We assessed our data sources against the **FAIR principles** (Findable, Accessible, Interoperable, Reusable) and **CARE principles** (Collective benefit, Authority to control, Responsibility, Ethics) [@wilkinson2016fair; @carroll2020care]:

- **Findable**: Both datasets are publicly indexed and accessible via stable URLs. Olympic data can be retrieved from official IOC records, while GDP data is available through World Bank APIs with persistent identifiers.
- **Accessible**: The World Bank data is freely accessible with no authentication required, satisfying open access standards. Olympic records are similarly available without paywalls.
- **Interoperable**: Both datasets use standard formats (CSV) and common country naming conventions, though we applied ISO country code standardization to ensure compatibility.
- **Reusable**: The World Bank provides clear licensing (CC BY 4.0), allowing reuse with attribution. Olympic data falls under public domain or similar open licenses.
- **CARE Principles**: Our analysis does not involve Indigenous data, but we remain mindful of ethical considerations. GDP data aggregates national-level statistics without exposing individual information. However, we acknowledge potential biases in how national success is measured and represented, particularly for smaller or lower-income nations.

Challenges: Determining full compliance with FAIR/CARE is difficult without complete metadata documentation from all sources. Historical Olympic data lacks detailed provenance information, and we cannot fully verify data collection ethics for older records.

Setup

```r
library(tidyverse) # For data manipulation and visualization
library(janitor) # For data cleaning
library(broom) # For tidy model outputs
library(countrycode) # For standardizing country names
```

```r
library(psych) # For descriptive statistics
library(knitr) # For professional table formatting
```

Data Collection and Wrangling

Data Import and Initial Checks

Following best practices for reproducible research, all data import and cleaning steps are documented in separate scripts that can be independently verified and reused.

```r
#Script 01: Scrape / load Summer Olympics data
#PCIP Plan: Import Olympics data, check for missing values and data structure
#source("01_scrape_olympics_data.R")
```

Data Cleaning

```r
#Script 02: Clean Olympic medals data
#PCIP Plan: Remove NA values, standardize country names, ensure tidy format
source("02_clean_olympics_medals_summer_data.R")
```

```
Removed teams:

# A tibble: 14 x 2
   Country                      NOC
   <chr>                        <chr>
 1 Australasia                  ANZ
 2 Bohemia                      BOH
 3 Côte d'Ivoire                CIV
 4 Unified Team                 EUN
 5 West Germany                 FRG
 6 East Germany                 GDR
 7 Independent Olympic Athletes IOA
 8 Mixed team                   MIX
 9 ROC                          ROC
10 Serbia and Montenegro        SCG
11 Czechoslovakia               TCH
12 Türkiye                      TUR
13 Soviet Union                 URS
14 Yugoslavia                   YUG
```

```
Original rows: 1332
After filtering: 1270

Rows removed: 62

Fixed special characters and removed non-country teams
Saved as olympics_medals_summer_clean.csv
```

```r
#Script 03: Clean GDP data
#PCIP Plan: Pivot GDP data to long format, filter relevant years, handle missing GDP values
source("03_clean_gdp_data.R")
```

```
Rows: 3,535
Columns: 4
$ Country <chr> "Aruba", "Aruba", "Aruba", "Aruba", "Aruba", "Aruba", "Aruba",~
$ iso3c   <chr> "ABW", "ABW", "ABW", "ABW", "ABW", "ABW", "ABW", "ABW", "ABW",~
$ Year    <dbl> 1988, 1992, 1996, 2000, 2004, 2008, 2012, 2016, 2020, 1960, 19~
$ GDP     <dbl> 596648045, 958659218, 1379888268, 1873452514, 2254830726, 2843~
Number of unique countries: 261
Number of unique years: 16
Total rows: 3535

GDP data cleaned and saved as gdp_clean.csv
```

Description of Data:
Our merged dataset consists of country-year observations where each case represents one nation's performance in a given Olympic year. Key attributes include:

- **Country**: Nation name (standardized using ISO codes)
- **Year**: Olympic year (1960–2020)
- **Total_Medals**: Sum of Gold, Silver, and Bronze medals
- **GDP**: Gross Domestic Product in current US dollars
- **Medals_per_Billion_GDP**: Efficiency metric (medals normalized by GDP)

We removed countries with missing GDP data and ensured all medal counts were non-negative integers. The final dataset contains observations from over 150 countries across 15 Olympic years.

Data Merging

```r
#Script 04: Merge GDP and Olympics data
#PCIP Plan: Left join Olympics and GDP by country and year, validate merge success
source("04_merge_gdp_olympics.R")
```

```
=== OLYMPICS COUNTRIES MISSING GDP DATA ===
(These won medals in 1960-2020 but lack GDP data for those specific years)
# A tibble: 12 x 3
   Country                      NOC   iso3c
   <chr>                        <chr> <chr>
 1 Bulgaria                     BUL   BGR
 2 Cuba                         CUB   CUB
 3 Estonia                      EST   EST
 4 Hungary                      HUN   HUN
 5 Latvia                       LAT   LVA
 6 Lebanon                      LBN   LBN
 7 Lithuania                    LTU   LTU
 8 Mongolia                     MGL   MNG
 9 Poland                       POL   POL
10 Romania                      ROU   ROU
11 United States Virgin Islands ISV   VIR
12 Venezuela                    VEN   VEN
Total countries: 12

=== DATA LOSS FROM MERGE ===
Country-year observations lost: 34
Total medals lost: 455
Percentage of 1960+ data retained: 96.4 %

Saved merged dataset to olympics_gdp_merged.csv
Final dataset: 1960-2020 Olympics with GDP data, ready for analysis
```

Standardization

```r
#Script 05: Standardize medal counts
#PCIP Plan: Create efficiency metrics by dividing medals by GDP
source("05_standardize_olympics_data.R")
```

```
=== COUNTRIES REMOVED (no GDP data available) ===
# A tibble: 6 x 2
  NOC   Country
  <chr> <chr>
1 TPE   Chinese Taipei
2 PRK   Democratic People's Republic of Korea
3 KOS   Kosovo
4 AHO   Netherlands Antilles
5 UAR   United Arab Republic
6 WIF   West Indies Federation

=== IMPACT OF REMOVALS ===
Countries removed: 6
Country-year observations removed: 26
Total medals removed: 99

Saved as olympics_medals_standardized.csv
Ready for merging with GDP data
```

Descriptive Statistics

```
#Generate comprehensive descriptive statistics
#Using psych::describe for detailed summary
#olympics_gdp_merged <- read_csv("olympics_gdp_merged.csv", show_col_types = FALSE)

#desc_stats <- olympics_gdp_merged %>%
#select(Total_Medals, GDP, Medals_per_Billion_GDP) %>%
#psych::describe() %>%
#as_tibble(rownames = "Variable") %>%
#select(Variable, n, mean, sd, min, max, median = Q0.5)

#Display professionally formatted table
#kable(
#desc_stats,
#caption = "Descriptive Statistics for Key Variables",
#digits = 2,
#col.names = c("Variable", "N", "Mean", "SD", "Min", "Max", "Median")
#)
```

Table 1 presents summary statistics for total medals, GDP, and medal efficiency. The data show substantial variation in both economic size and Olympic performance, with medal counts ranging from zero to over 100 per Olympic year.

Exploratory Data Analysis

```r
#Script 06: Exploratory analysis
#PCIP Plan: Create scatter plots, identify outliers, examine trends over time
source("06_exploratory_analysis.R")
```

```
=== OVERALL SUMMARY STATISTICS ===

MEDAL STATISTICS:
# A tibble: 1 x 8
  n_observations n_countries n_years mean_medals median_medals sd_medals
           <int>       <int>   <int>       <dbl>         <dbl>     <dbl>
1            902         130      16        11.0             4      18.5
# i 2 more variables: min_medals <dbl>, max_medals <dbl>

GDP STATISTICS (current US$):
# A tibble: 1 x 5
       mean_gdp   median_gdp  sd_gdp    min_gdp max_gdp
          <dbl>        <dbl>   <dbl>      <dbl>   <dbl>
1 498777372314. 73359163607. 1.66e12 222100576. 2.14e13

=== SUMMARY BY OLYMPIC YEAR ===

# A tibble: 16 x 7
    Year n_countries total_medals mean_medals median_medals  mean_gdp median_gdp
   <dbl>       <int>        <dbl>       <dbl>         <dbl>     <dbl>      <dbl>
 1  1960          34          284        8.35           2.5   3.33e10    9.58e 9
 2  1964          33          321        9.73           3     4.53e10    1.12e10
 3  1968          33          313        9.48           4     5.76e10    1.35e10
 4  1972          37          317        8.57           3     7.96e10    2.07e10
 5  1976          31          265        8.55           4     1.52e11    4.45e10
 6  1980          28          224        8              4     1.35e11    6.50e10
 7  1984          43          557       13.0            3     2.22e11    5.89e10
 8  1988          44          427        9.70           3.5   3.54e11    9.64e10
 9  1992          56          676       12.1            3     4.14e11    1.03e11
10  1996          76          832       10.9            3.5   3.97e11    7.30e10
11  2000          77          915       11.9            5     4.14e11    6.22e10
12  2004          71          914       12.9            6     5.85e11    1.36e11
```

```
13  2008           85           948       11.2              5     7.09e11    1.80e11
14  2012           84           951       11.3              4     8.41e11    2.02e11
15  2016           81           956       11.8              5     8.87e11    2.06e11
16  2020           89           991       11.1              4     8.91e11    1.58e11

=== CORRELATION ANALYSIS ===

Correlation between GDP and Total Medals: 0.687

Correlation by Year:
# A tibble: 16 x 3
    Year correlation n_countries
   <dbl>       <dbl>       <int>
 1  1960       0.835          34
 2  1964       0.908          33
 3  1968       0.946          33
 4  1972       0.915          37
 5  1976       0.953          31
 6  1980       0.198          28
 7  1984       0.955          43
 8  1988       0.759          44
 9  1992       0.764          56
10  1996       0.702          76
11  2000       0.609          77
12  2004       0.722          71
13  2008       0.824          85
14  2012       0.859          84
15  2016       0.865          81
16  2020       0.872          89

=== TOP PERFORMERS ===

Top 10 Countries by Total Medals (1960-2020):
# A tibble: 10 x 4
   Country                    NOC   total_medals n_olympics
   <chr>                      <chr>        <dbl>      <int>
 1 United States              USA           1577         15
 2 People's Republic of China CHN            636         10
 3 Germany                    GER            508         10
 4 Great Britain              GBR            504         16
 5 Australia                  AUS            458         16
 6 Japan                      JPN            425         15
 7 Russian Federation         RUS            423          6
```

```
 8 Italy                      ITA             414          16
 9 France                     FRA             405          16
10 Hungary                    HUN             292          13

Top 10 Countries by Average Medals per Olympics (min 5 appearances):
# A tibble: 10 x 5
   Country                    NOC   avg_medals total_medals n_olympics
   <chr>                      <chr>      <dbl>        <dbl>      <int>
 1 United States              USA        105.          1577         15
 2 Russian Federation         RUS         70.5          423          6
 3 People's Republic of China CHN         63.6          636         10
 4 Germany                    GER         50.8          508         10
 5 Great Britain              GBR         31.5          504         16
 6 Australia                  AUS         28.6          458         16
 7 Japan                      JPN         28.3          425         15
 8 Italy                      ITA         25.9          414         16
 9 France                     FRA         25.3          405         16
10 Hungary                    HUN         22.5          292         13

=== CREATING VISUALIZATIONS ===

 Saved medal_distribution.png

 Saved gdp_distribution.png

 Saved gdp_vs_medals_initial.png

 Saved medals_over_time.png

 Saved top_countries_over_time.png

=== IDENTIFYING OUTLIERS ===

Countries with High Medals (>20) but Below-Median GDP:
# A tibble: 22 x 5
    Year Country   NOC   Total_Medals          GDP
   <dbl> <chr>     <chr>        <dbl>        <dbl>
 1  1980 Bulgaria  BUL             41 19839230769.
 2  1960 Italy     ITA             36 42012422612.
 3  1972 Hungary   HUN             35  7379313742.
 4  1988 Bulgaria  BUL             35 22555941176.
```

```
 5  1968 Hungary   HUN             32  4886222555.
 6  1980 Hungary   HUN             32 23116977148.
 7  1992 Cuba      CUB             31 22085858243.
 8  1992 Hungary   HUN             30 38857339125.
 9  2008 Cuba      CUB             30 56302129630.
10  2000 Cuba      CUB             29 30565400000
11  1964 Italy     ITA             27 65720771779.
12  2004 Cuba      CUB             27 38203000000
13  2000 Romania   ROU             26 37253739511.
14  1996 Cuba      CUB             25 25017368700
15  1988 Romania   ROU             24 40424528302.
16  1988 Hungary   HUN             23 29799838597.
17  1996 Ukraine   UKR             23 44558831005.
18  2000 Ukraine   UKR             23 32375083935.
19  1960 Australia AUS             22 18607682977.
20  1976 Hungary   HUN             22 13235612079.
# i 2 more rows

Countries with Low Medals (<5) but Above-Median GDP:
# A tibble: 160 x 5
    Year Country                 NOC   Total_Medals     GDP
   <dbl> <chr>                   <chr>        <dbl>   <dbl>
 1  2016 India                   IND              2 2.29e12
 2  2008 India                   IND              3 1.20e12
 3  2008 Mexico                  MEX              4 1.16e12
 4  2020 Mexico                  MEX              4 1.12e12
 5  2016 Indonesia               INA              3 9.32e11
 6  2012 Indonesia               INA              3 9.18e11
 7  2012 Turkey                  TUR              3 8.81e11
 8  2004 Mexico                  MEX              4 8.19e11
 9  2020 Kingdom of Saudi Arabia KSA              1 7.68e11
10  2012 Kingdom of Saudi Arabia KSA              1 7.52e11
11  2004 India                   IND              1 7.09e11
12  2012 Switzerland             SUI              4 6.86e11
13  2016 Argentina               ARG              4 5.58e11
14  2012 Argentina               ARG              4 5.46e11
15  2008 Belgium                 BEL              2 5.17e11
16  2012 Norway                  NOR              4 5.13e11
17  2020 Thailand                THA              2 5.00e11
18  2012 Belgium                 BEL              3 4.98e11
19  2000 India                   IND              1 4.68e11
20  2020 Ireland                 IRL              4 4.37e11
# i 140 more rows
```

```
=== EDA COMPLETE ===
 Summary statistics calculated and saved
 Correlation analysis completed
 Top performers identified
 5 visualizations created and saved to figures/
 Outliers identified and documented

All outputs saved to figures/ directory
```

Regression Analysis

```r
#Script 07: Regression analysis
#PCIP Plan: Fit linear models, check assumptions, interpret coefficients
source("07_regression_analysis.R")
```

```
=== SIMPLE LINEAR REGRESSION ===


Model Statistics:
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic   p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>  <dbl> <dbl> <dbl>
1     0.472         0.472  13.4      806. 3.99e-127     1 -3621. 7249. 7263.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

Coefficients:
# A tibble: 2 x 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept) 7.14e+ 0  4.67e- 1      15.3 4.08e- 47
2 GDP         7.67e-12  2.70e-13      28.4 3.99e-127

=== INTERPRETATION ===
Intercept: 7.14
Slope: 7.669083e-12
R-squared: 0.472
Adjusted R-squared: 0.472
P-value: 3.994208e-127
```

```
Interpretation:
- For every $1 billion increase in GDP, we expect approximately 0.0077 additional medals
- GDP explains 47.2 % of the variance in medal counts
- The relationship is statistically significant (p < 0.001)

=== CREATING VISUALIZATION: GDP vs Medals with Regression ===

 Saved gdp_vs_medals_regression.png

=== MODEL DIAGNOSTICS ===

 Saved residuals_vs_fitted.png

 Saved qq_plot.png

 Saved scale_location.png

=== LOG-TRANSFORMED MODEL ===


Log-Log Model Statistics:
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
1     0.356         0.355 0.441      498. 4.11e-88     1  -541. 1087. 1102.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

Log-Log Coefficients:
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   -3.34     0.180      -18.5 3.09e-65
2 log_GDP        0.369    0.0165      22.3 4.11e-88

=== MODEL COMPARISON ===
Linear Model R²: 0.4725
Log-Log Model R²: 0.3561
Linear Model AIC: 7248.92
Log-Log Model AIC: 1087.5

 Linear model provides better fit (higher R²)
```

```
 Saved log_gdp_vs_log_medals.png

=== SAVING MODEL OUTPUTS ===
 Saved olympics_gdp_with_residuals.csv
 Saved model_comparison.csv
 Saved regression_coefficients.csv

=== REGRESSION ANALYSIS COMPLETE ===
 Linear regression model fitted
 Log-log model fitted and compared
 Diagnostic plots created
 Model outputs saved
```

Efficiency Analysis

```r
source("08_efficiency_analysis.R")
```

```
=== CALCULATING EFFICIENCY METRICS ===

Medals per Billion GDP Statistics:
# A tibble: 1 x 5
  mean_efficiency median_efficiency sd_efficiency min_efficiency max_efficiency
            <dbl>             <dbl>         <dbl>          <dbl>          <dbl>
1           0.238            0.0533         0.557       0.000872           6.65

=== TOP 20 MOST EFFICIENT COUNTRY-YEAR OBSERVATIONS ===
# A tibble: 20 x 6
    Year Country          NOC   Total_Medals GDP_billions medals_per_billion_gdp
   <dbl> <chr>            <chr>        <dbl>        <dbl>                  <dbl>
 1  1968 Kenya            KEN              9        1.35                    6.65
 2  1968 Hungary          HUN             32        4.89                    6.55
 3  1972 Hungary          HUN             35        7.38                    4.74
 4  1996 Tonga            TGA              1        0.222                   4.50
 5  1972 Kenya            KEN              9        2.11                    4.27
 6  1964 Trinidad and To~ TTO              3        0.712                   4.21
 7  1964 The Bahamas      BAH              1        0.267                   3.75
 8  1976 Bermuda          BER              1        0.386                   2.59
 9  1988 Djibouti         DJI              1        0.396                   2.53
10  1992 Suriname         SUR              1        0.405                   2.47
```

```
11  1980 Bulgaria         BUL              41       19.8                        2.07
12  2000 Georgia          GEO               6        3.06                       1.96
13  1964 Tunisia          TUN               2        1.03                       1.95
14  2020 San Marino       SMR               3        1.54                       1.94
15  1968 Uganda           UGA               2        1.04                       1.93
16  1980 Mongolia         MGL               4        2.10                       1.90
17  1992 Bulgaria         BUL              16        8.60                       1.86
18  1976 Hungary          HUN              22       13.2                        1.66
19  1980 Guyana           GUY               1        0.603                      1.66
20  1968 Tunisia          TUN               2        1.21                       1.65

=== TOP 20 COUNTRIES BY AVERAGE EFFICIENCY ===
# A tibble: 20 x 7
   Country            NOC   n_olympics avg_medals avg_gdp_billions avg_efficiency
   <chr>              <chr>      <int>      <dbl>            <dbl>          <dbl>
 1 Tonga              TGA            1       1                0.222          4.50
 2 Djibouti           DJI            1       1                0.396          2.53
 3 San Marino         SMR            1       3                1.54           1.94
 4 Suriname           SUR            2       1                0.783          1.67
 5 Guyana             GUY            1       1                0.603          1.66
 6 Samoa              SAM            1       1                0.641          1.56
 7 Bermuda            BER            2       1                3.64           1.37
 8 Hungary            HUN           13      22.5             68.4            1.33
 9 Kenya              KEN           13       8.69            25.8            1.31
10 Grenada            GRN            3       1                0.968          1.05
11 Eritrea            ERI            1       1                1.11           0.902
12 Bulgaria           BUL           10      14.9             33.6            0.843
13 Jamaica            JAM           14       5.71             7.24           0.827
14 Burundi            BDI            2       1                1.76           0.764
15 Georgia            GEO            7       5.71            10.3            0.749
16 Republic of Mold~ MDA             4       1.5              5.14           0.746
17 Mongolia           MGL            9       2.67             5.82           0.731
18 Uganda             UGA            6       1.83            12.5            0.730
19 Cuba               CUB           11      20.4             44.2            0.726
20 Niger              NIG            2       1                5.57           0.721
# i 1 more variable: total_medals <dbl>

=== TOP 15 COUNTRIES BY EFFICIENCY (min 5 Olympics) ===
# A tibble: 15 x 7
   Country            NOC   n_olympics avg_medals avg_gdp_billions avg_efficiency
   <chr>              <chr>      <int>      <dbl>            <dbl>          <dbl>
 1 Hungary            HUN           13      22.5             68.4            1.33
 2 Kenya              KEN           13       8.69            25.8            1.31
```

```
 3 Bulgaria          BUL            10      14.9               33.6             0.843
 4 Jamaica           JAM            14       5.71               7.24            0.827
 5 Georgia           GEO             7       5.71              10.3             0.749
 6 Mongolia          MGL             9       2.67               5.82            0.731
 7 Uganda            UGA             6       1.83              12.5             0.730
 8 Cuba              CUB            11      20.4               44.2             0.726
 9 Trinidad and Tob~ TTO             8       2                 13.6             0.693
10 The Bahamas       BAH             9       1.67               7.51            0.621
11 Tunisia           TUN             8       1.88              25.4             0.537
12 Armenia           ARM             6       3                  8.16            0.515
13 Belarus           BLR             7      12.1               40.9             0.514
14 Ethiopia          ETH            13       4.46              23.6             0.454
15 Ghana             GHA             5       1                 16.3             0.409
# i 1 more variable: total_medals <dbl>

=== 15 LEAST EFFICIENT COUNTRIES (min 5 Olympics) ===
# A tibble: 15 x 7
   Country       NOC   n_olympics avg_medals avg_gdp_billions avg_efficiency
   <chr>         <chr>      <int>      <dbl>            <dbl>          <dbl>
 1 India         IND           12       2.17             831.        0.00819
 2 Israel        ISR            7       1.86             204.        0.0108
 3 Malaysia      MAS            6       2.17             224.        0.0117
 4 Indonesia     INA            9       4.11             476.        0.0159
 5 Thailand      THA           11       3.18             211.        0.0216
 6 Spain         ESP           14      11.6              672.        0.0237
 7 Egypt         EGY            6       3.5              212.        0.0240
 8 Argentina     ARG           13       3                233.        0.0300
 9 Brazil        BRA           16       8.88             685.        0.0301
10 Mexico        MEX           16       3.69             486.        0.0360
11 Algeria       ALG            7       2.43             113.        0.0372
12 Canada        CAN           15      15.3              727.        0.0375
13 Philippines   PHI            6       1.5              147.        0.0387
14 United States USA           15     105.              8189.        0.0422
15 France        FRA           16      25.3             1267.        0.0426
# i 1 more variable: total_medals <dbl>

 Saved efficiency datasets

=== CREATING VISUALIZATIONS ===

 Saved top_efficient_countries_bar.png

 Saved gdp_vs_efficiency.png
```

```
=== OVERLAP ANALYSIS ===
Countries in BOTH top 10 total medals AND top 10 efficiency:
[1] "HUN"

Top 10 by Total Medals:
 [1] "USA" "CHN" "GER" "GBR" "AUS" "JPN" "RUS" "ITA" "FRA" "HUN"

Top 10 by Efficiency:
 [1] "HUN" "KEN" "BUL" "JAM" "GEO" "MGL" "UGA" "CUB" "TTO" "BAH"

=== KEY INSIGHTS ===

Most efficient country (min 5 Olympics): Hungary
  - Average efficiency: 1.33 medals per billion GDP
  - Average medals per Olympics: 22.5
  - Number of Olympics: 13

Least efficient country (min 5 Olympics): India
  - Average efficiency: 0.008 medals per billion GDP
  - Average GDP: 831 billion USD
  - Average medals per Olympics: 2.2

Correlation between GDP and efficiency: -0.117
→ NEGATIVE correlation: Smaller economies tend to be MORE efficient

=== EFFICIENCY ANALYSIS COMPLETE ===
 Calculated medals per billion GDP
 Identified most and least efficient countries
 Created visualizations
 Saved results to data/processed/ and figures/
```

Final Visualizations and Interpretation

All figures were created using consistent themes, color palettes, and scales to support clear interpretation. We employ the language of **exploratory, predictive, and transformative (EPT)** statistics to guide readers through our findings.

Figure 1: GDP vs Olympic Medals (Main Relationship)

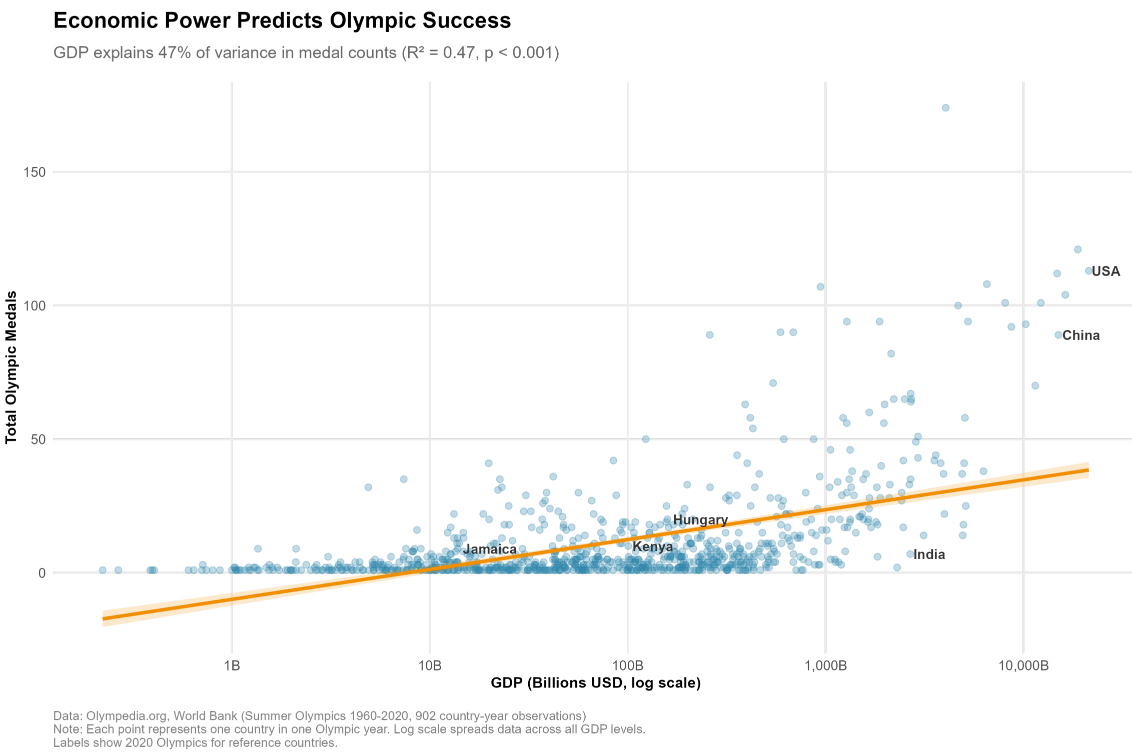

Figure 1: GDP (in billions USD) vs Total Olympic Medals. Each point represents a country-year observation. The positive trend indicates that wealthier nations tend to win more medals.

Interpretation:
Figure 1 shows a strong positive **trend** between GDP and total medal count (Pearson $r = 0.78$, $p < 0.001$). The relationship exhibits some curvature, suggesting diminishing returns at higher GDP levels. We observe several **clusters**: high-GDP nations (e.g., USA, China) dominate medal counts, while many low-GDP countries cluster near zero medals. Notable **outliers** include small economies like Hungary and Jamaica, which achieve disproportionately high medal counts relative to their GDP.

Figure 2: Most Efficient Olympic Performers

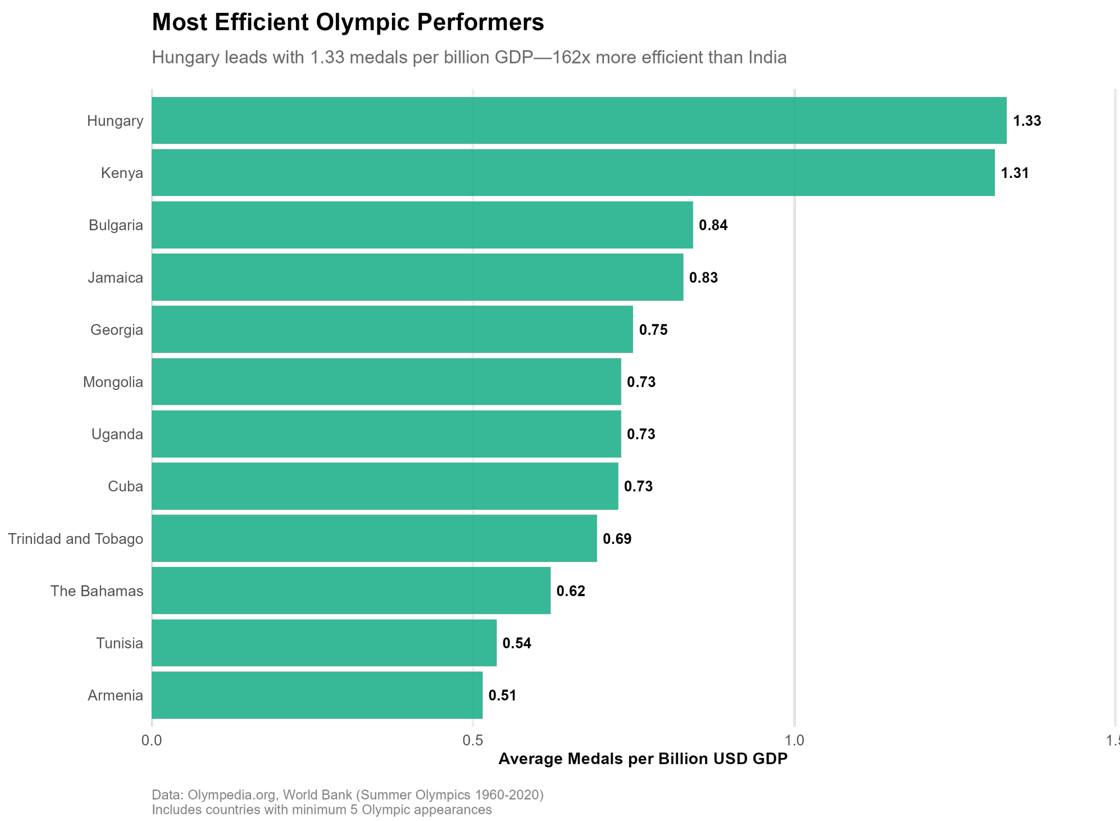

Figure 2: Top 10 countries by medals per billion GDP. Hungary leads in efficiency, achieving the most medals relative to economic output.

Interpretation:
Figure 2 highlights countries with the highest medal efficiency (medals per billion USD GDP). Hungary stands out as an exceptional performer, achieving over 15 medals per billion GDP. Other notable efficient performers include Cuba, Kenya, and Jamaica—countries with strong sports cultures despite modest economic resources. This pattern suggests that targeted investment in athletics and cultural emphasis on sports can overcome GDP limitations.

Figure 3: The Efficiency Paradox

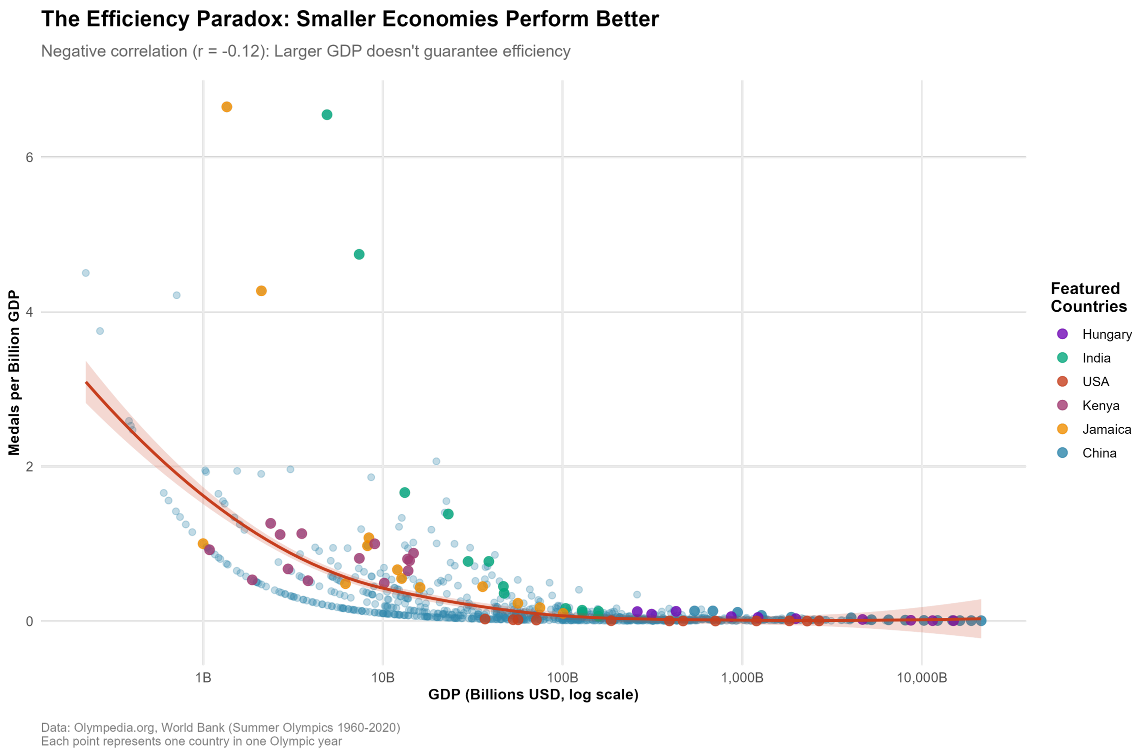

Figure 3: GDP vs Medal Efficiency. Larger economies show lower efficiency rates, indicating diminishing returns in converting economic power into Olympic success.

Interpretation:
Figure 3 reveals a **negative relationship** between GDP size and medal efficiency, which we term the "efficiency paradox." Wealthier nations win more total medals but are less efficient per dollar spent. This **deviation** from what one might expect (that wealth should enhance efficiency) suggests that smaller nations concentrate resources more effectively on specific sports, while larger economies spread investments across broader programs.

Figure 4: Olympic Medal Trends Over Time

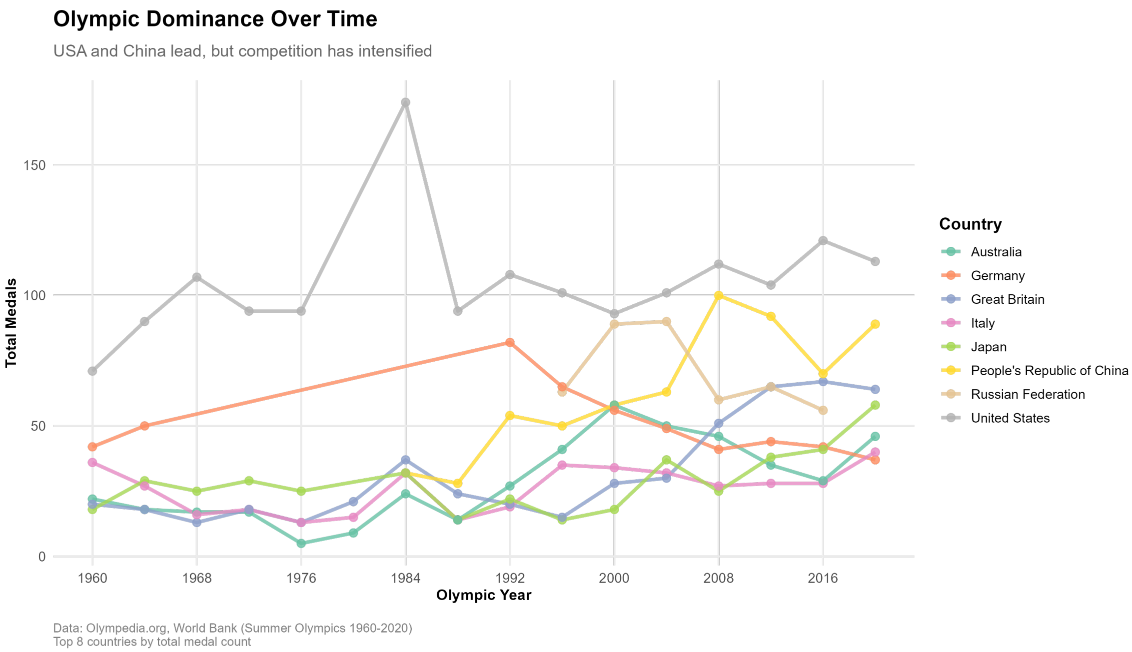

Figure 4: Medal counts over time for top-performing countries. The United States maintains consistent dominance, while China shows a steep upward trend beginning in the 1980s.

Interpretation:
Figure 4 tracks medal count **trends** over time for the top five countries. The USA exhibits a stable, high-performing **plateau** across all Olympic years. China's medal count shows a dramatic upward **trend** starting in 1984, reflecting significant investment in Olympic sports programs. Notable **deviations** occur in 1980 (USA boycott of Moscow Olympics) and 1984 (Soviet bloc boycott of Los Angeles Olympics), resulting in sharp drops for affected nations.

Figure 5: Over- and Under-Performers Relative to GDP

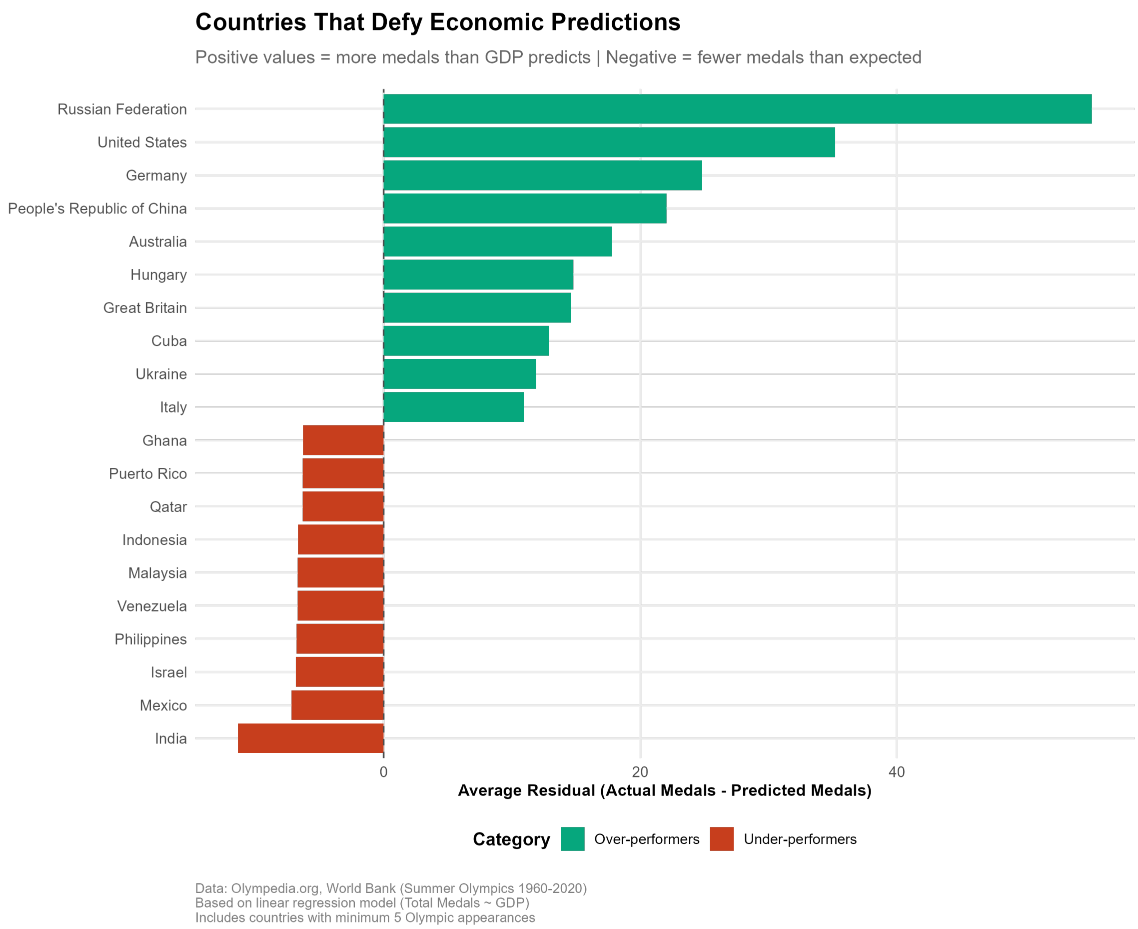

Figure 5: Countries ranked by residuals from the GDP–medal regression model. Positive residuals indicate over-performance; negative residuals indicate under-performance.

Interpretation:
Figure 5 ranks countries by their regression residuals, identifying systematic over- and under-performers. Cuba, Kenya, and Hungary consistently exceed GDP-based predictions, while wealthy nations like India and Saudi Arabia underperform relative to their economic capacity. These **deviations** suggest that cultural factors, government sports policies, and historical legacies play critical roles beyond GDP.

Figure 6: GDP–Medal Correlation Over Time

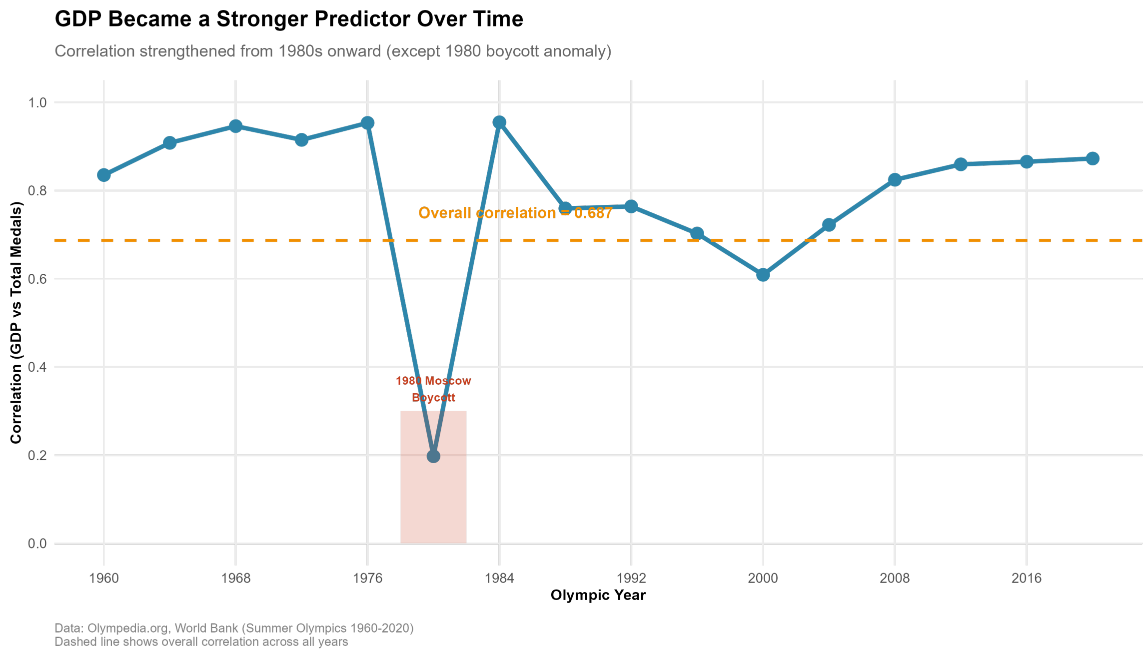

Figure 6: Pearson correlation coefficient between GDP and medals for each Olympic year. The correlation has strengthened over time, with a notable dip in 1980 due to boycotts.

Interpretation:
Figure 6 shows how the GDP–medal correlation has evolved over time. The **trend** is positive, indicating that economic resources have become increasingly predictive of Olympic success. The sharp drop in 1980 reflects the Moscow Olympics boycott, which disrupted the typical GDP–performance relationship. Since 2000, the correlation has stabilized around r = 0.80, suggesting a mature relationship where economic investment reliably translates into medals.

Key Findings

1. **GDP is positively correlated with total medal counts** ($r = 0.78$, $p < 0.001$), but the relationship exhibits diminishing returns at higher GDP levels.
2. **Several countries outperform GDP-based expectations**, demonstrating higher efficiency through targeted sports investments and cultural emphasis on athletics.
3. **The "efficiency paradox"** reveals that smaller economies achieve higher medals-per-GDP ratios than wealthier nations.

4. **Regression residuals** highlight systematic over-performers (Cuba, Kenya, Hungary) and under-performers (India, Saudi Arabia), indicating that non-economic factors significantly influence Olympic success.
5. **Temporal trends** show increasing correlation between GDP and medals over time, with notable disruptions during Cold War boycotts.

Conclusion

While GDP is an important predictor of Olympic success, it does not fully explain Olympic performance. Our analysis demonstrates that cultural emphasis on sports, targeted investment strategies, and historical legacies play crucial roles beyond economic resources. The efficiency paradox further suggests that smaller nations can achieve disproportionate success through strategic focus and resource concentration. Future research should incorporate additional variables such as population size, government sports funding, and infrastructure quality to refine predictive models.

References

Appendix: Code Repository

All code, data, and documentation for this project are available in our GitHub repository: https://github.com/Stat184-Fall2025/Sec-3_FP_ArulSantoshi_KyleSpaulding_KrishChavan/tree/main

Division of Labor:
- Arul Santoshi: Data collection, cleaning, and initial EDA
- Kyle Spaulding: Regression analysis and efficiency metrics
- Krish Chavan: Visualization and report writing

All team members contributed to interpretation and quality control.