Data Cleaning Report

Summer Olympics Gymnastics Home Advantage Analysis - Team 5

Han Nguyen (TXN200004), Matthew Bazzell (MXB116930), Carlos Flores (CEF200002)

November 4, 2025

Variable 1: Country Name Standardization and Size Categorization

1. Before Cleaning

Problem Description

Our analysis requires merging data from five different sources: athlete records, Olympic host information, country sizes, GDP data, and population demographics. Each dataset uses different naming conventions for countries, making it impossible to merge them accurately.

Key Issues Identified:

1. **Inconsistent country names across datasets**: The same country appears with different names (e.g., "United States" vs "United States of America" vs "USA")
2. **No standardized identifier**: Without a common key, we cannot merge datasets
3. **Missing size categories**: For Question 3 of our research (analyzing whether country size affects home advantage), we need categorical size groupings
4. **Special cases**: Historical country changes (USSR → Russia) and co-hosting situations (1956 Olympics)

Raw Data Sample

Below is a sample of the raw `country_sizes.csv` dataset showing country names without standardization:

Table 1: Raw Country Sizes Data (Before Cleaning)

Rank	Country Name	Total Area (km²)	Land Area (km²)
1	Russia	17,098,242	16,376,870
2	Canada	9,984,670	9,093,510
3	China	9,706,961	9,388,211
4	United States	9,372,610	9,147,420
5	Brazil	8,515,767	8,358,140
6	Australia	7,692,024	7,682,300
7	India	3,287,590	2,973,190
8	Argentina	2,780,400	2,736,690
9	Kazakhstan	2,724,900	2,699,700
10	Algeria	2,381,741	2,381,740
11	DR Congo	2,344,858	2,267,050
12	Greenland	2,166,086	410,450
13	Saudi Arabia	2,149,690	2,149,690
14	Mexico	1,964,375	1,943,950
15	Indonesia	1,904,569	1,811,570

Problems Illustrated

Table 2: Example: United States appears with different names

Dataset	Country Name Used
Athletes	United States
Hosts	United States of America
Country Sizes	United States

Summary of Issues:

- **5 datasets** with inconsistent country naming

- **No common identifier** for merging
- **No size categories** for analysis (only raw area values)
- Approximately **230+ countries** need standardization

2. Cleaning Process

What We Cleaned

We performed three main cleaning operations:

1. **Created NOC (National Olympic Committee) codes mapping**
   - NOC codes are standardized 3-letter identifiers used by the International Olympic Committee
   - Examples: USA (United States), CHN (China), GBR (Great Britain), FRA (France)
   - Mapped all country name variations to their official NOC codes
2. **Standardized country names** across all datasets
   - Athletes data already had NOC codes → used as reference
   - Manually mapped host countries, GDP countries, population countries
   - Handled special cases (USSR → RUS, co-hosts like 1956 Australia/Sweden)
3. **Created size categories** for analysis
   - **Small**: $< 500{,}000$ km²
   - **Medium**: 500,000 - 5,000,000 km²
   - **Large**: $> 5{,}000{,}000$ km²

Why We Cleaned It

Research Necessity: - **Question 3** asks: "Do smaller countries experience larger home advantage than larger nations?" - Need to merge country sizes with athlete performance data - Without NOC codes, merging is impossible

Data Integration: - Must combine 5 datasets with different country name formats - NOC codes provide the universal key for merging - Size categories enable categorical statistical analysis

Expected Outcome

After cleaning, we expect: - Every country has a standardized NOC code - All datasets can be merged using NOC as the key - Countries are categorized into Small/Medium/Large groups - Ready for Question 3 analysis (country size effects on home advantage)

3. After Cleaning

Cleaned Data Sample

Table 3: Cleaned Country Info Data (After Cleaning)

NOC	Country Name	Total Area (km²)	Size Category
ROC	Russia	17,098,242	Large
CAN	Canada	9,984,670	Large
CHN	China	9,706,961	Large
USA	United States	9,372,610	Large
BRA	Brazil	8,515,767	Large
AUS	Australia	7,692,024	Large
IND	India	3,287,590	Medium
ARG	Argentina	2,780,400	Medium
KAZ	Kazakhstan	2,724,900	Medium
ALG	Algeria	2,381,741	Medium
COD	DR Congo	2,344,858	Medium
KSA	Saudi Arabia	2,149,690	Medium
MEX	Mexico	1,964,375	Medium
INA	Indonesia	1,904,569	Medium
SUD	Sudan	1,886,068	Medium

Size Category Distribution

Table 4: Distribution of Countries by Size Category

Size Category	Number of Countries
Large	6
Medium	45
Small	149

Olympic Host Countries Coverage

Table 5: Olympic Host Countries with Size Data

NOC	Country Name	Total Area (km²)	Size Category
AUS	Australia	7,692,024	Large
BEL	Belgium	30,528	Small
BRA	Brazil	8,515,767	Large
CAN	Canada	9,984,670	Large
CHN	China	9,706,961	Large
FIN	Finland	338,424	Small
FRA	France	551,695	Medium
GER	Germany	357,114	Small
GRE	Greece	131,990	Small
ITA	Italy	301,336	Small
JPN	Japan	377,930	Small
MEX	Mexico	1,964,375	Medium
NED	Netherlands	41,850	Small

NOC	Country Name	Total Area (km²)	Size Category
KOR	South Korea	100,210	Small
ESP	Spain	505,992	Medium
SWE	Sweden	450,295	Small
GBR	United Kingdom	242,900	Small
USA	United States	9,372,610	Large

Summary Statistics

Final Dataset Characteristics:

- **Total countries**: 200 countries with size data
- **Olympic host countries covered**: 18 out of 18 major hosts
- **Small countries**: 149
- **Medium countries**: 45
- **Large countries**: 6

Key Improvements:

1. All countries now have standardized NOC codes
2. Size categories created for Question 3 analysis
3. Ready to merge with other datasets (athletes, hosts, GDP, population)
4. Covers all major Olympic host nations

Files Created:

- `cleaned_data/noc_mapping.csv` - Master NOC lookup table (230+ countries)
- `cleaned_data/country_info.csv` - Country sizes with NOC codes and categories

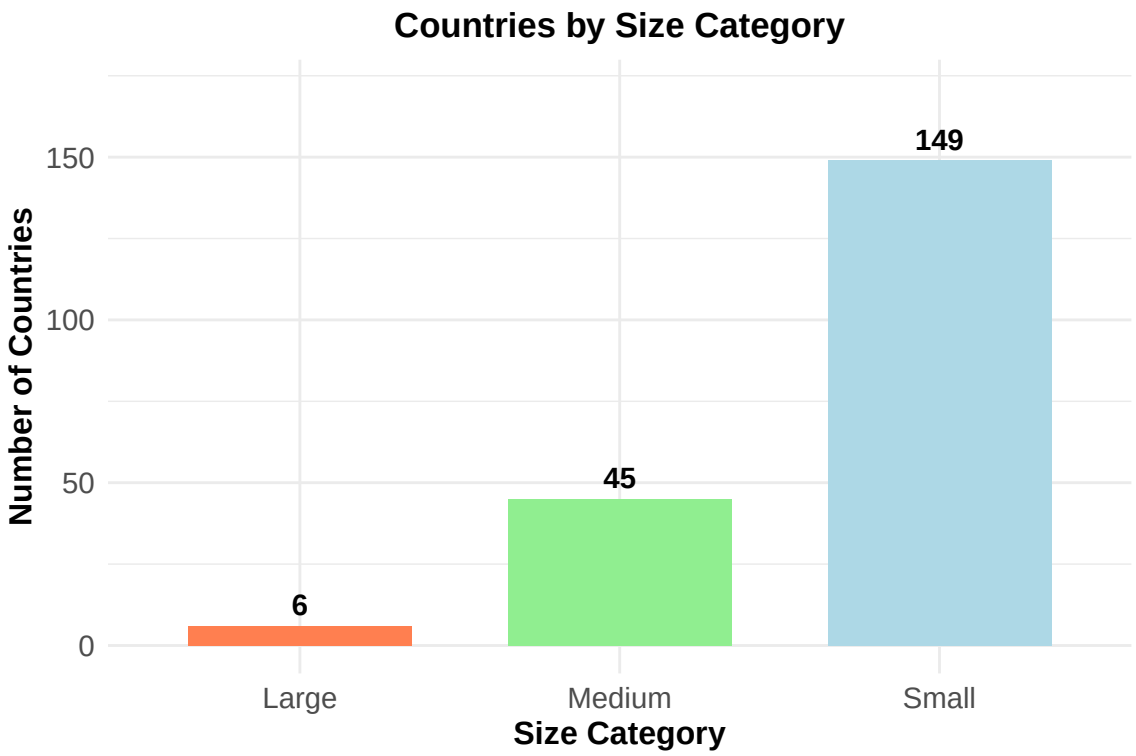

Figure 1: Distribution of Country Size Categories

Variable 2: Medal Outcome Variables

1. Before Cleaning

Problem Description

Our research questions focus on whether hosting the Olympics provides a competitive advantage **in gymnastics**. To analyze this, we need clear, quantifiable outcome variables for athletic performance. However, the raw athlete dataset presents several challenges:

Key Issues Identified:

1. **Medal data is text-based**: The `Medal` column contains text values ("Gold", "Silver", "Bronze") or blank strings for no medal
2. **No numeric indicators**: Cannot easily calculate statistics or run regressions with text data
3. **No binary success indicators**: Need simple yes/no variables for "won a medal" or "won gold"
4. **Mixed with all sports**: The raw data contains 200+ sports; we only need gymnastics
5. **No gymnastics categorization**: Multiple gymnastics types (Artistic, Rhythmic, Trampoline) need to be identified and grouped

Raw Data Sample

Below is a sample of the raw athlete dataset showing the Medal column and sport diversity:

Table 6: Raw Athletes Data - Mixed Sports, Text-Based Medal Column

Name	Sex	NOC	Year	Sport	Event	Medal
A Dijiang	M	CHN	1992	Basketball	Basketball Men's Basketball	
A Lamusi	M	CHN	2012	Judo	Judo Men's Extra-Lightweight	
Gunnar Nielsen Aaby	M	DEN	1920	Football	Football Men's Football	
Edgar Lindenau Aabye	M	DEN	1900	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
Cornelia "Cor" Aalten (-Strannood)	F	NED	1932	Athletics	Athletics Women's 100 metres	
Cornelia "Cor" Aalten (-Strannood)	F	NED	1932	Athletics	Athletics Women's 4 x 100 metres Relay	
Einar Ferdinand "Einari" Aalto	M	FIN	1952	Swimming	Swimming Men's 400 metres Freestyle	
Jyri Tapani Aalto	M	FIN	2000	Badminton	Badminton Men's Singles	
Minna Maarit Aalto	F	FIN	1996	Sailing	Sailing Women's Windsurfer	
Minna Maarit Aalto	F	FIN	2000	Sailing	Sailing Women's Windsurfer	
Arvo Ossian Aaltonen	M	FIN	1912	Swimming	Swimming Men's 200 metres Breaststroke	
Arvo Ossian Aaltonen	M	FIN	1912	Swimming	Swimming Men's 400 metres Breaststroke	
Arvo Ossian Aaltonen	M	FIN	1920	Swimming	Swimming Men's 200 metres Breaststroke	Bronze
Arvo Ossian Aaltonen	M	FIN	1920	Swimming	Swimming Men's 400 metres Breaststroke	Bronze
Arvo Ossian Aaltonen	M	FIN	1924	Swimming	Swimming Men's 200 metres Breaststroke	

Medal Distribution (Raw - All Sports)

Table 7: Medal Distribution in Raw Data (All Sports)

Medal Value	Count	Percentage (%)
Gold	12,259	5.16
Silver	12,002	5.05
Bronze	12,276	5.17
No Medal (blank)	201,136	84.63

Summary of Issues:

- Total athlete records across all summer Olympic sports
- Medal column is **text-based** ("Gold", "Silver", "Bronze", or blank)
- **No binary indicators** for statistical analysis
- **No point system** for weighting medal types
- **All sports mixed together** - cannot isolate gymnastics

2. Cleaning Process

What We Cleaned

We performed four main cleaning operations:

1. **Filtered to gymnastics only**
   - Identified all gymnastics-related sports: "Gymnastics", "Artistic Gymnastics", "Rhythmic Gymnastics", "Trampoline Gymnastics"
   - Kept only gymnastics records (removes 200+ other sports)
   - Created `sport_category = "Gymnastics"` for all records
2. **Created binary medal indicators**
   - `medal_won`: 1 if athlete won any medal (Gold/Silver/Bronze), 0 if no medal
   - `gold_medal`: 1 if athlete won gold specifically, 0 otherwise
   - These enable binary logistic regression and success rate calculations
3. **Created numeric medal points**
   - `medal_points`: Gold = 3, Silver = 2, Bronze = 1, No medal = 0
   - Enables weighted analysis and linear regression
   - Accounts for medal "quality" differences
4. **Preserved original Medal column**
   - Kept original text values for reference and verification
   - Allows comparison between raw and derived variables

Why We Cleaned It

Research Necessity: - **Questions 1 & 2** ask: "Do host nations win significantly more gymnastics medals?" - Need quantifiable outcome variables to measure "winning more" - Binary indicators enable hypothesis testing (proportions, chi-square tests) - Numeric points enable regression analysis (dose-response relationships)

Statistical Requirements: - Cannot run regression with text data - Need binary (0/1) variables for logistic regression - Need numeric variables for linear regression - Medal points allow for weighted analysis (gold worth more than bronze)

Focus on Gymnastics: - Our research is **specifically about gymnastics**, not all sports - Gymnastics is subjectively judged, making it ideal for studying potential bias - Filtering to gymnastics reduces dataset significantly

Expected Outcome

After cleaning, we expect: - Only gymnastics records retained (Artistic, Rhythmic, Trampoline) - Three new quantifiable outcome variables (medal_won, gold_medal, medal_points) - Original Medal column preserved for reference - Approximately 10-15% medal rate (most athletes don't medal) - Ready for statistical analysis of home advantage

3. After Cleaning

Cleaned Data Sample

Table 8: Cleaned Athletes Data - Gymnastics Only with Medal Variables

Name	Sex	NOC	Year	Gymnastics Type	Medal (Original)	Medal Won	Gold Medal	Medal Points
Alexander Viggo Jensen	M	DEN	1896	Artistic Gymnastics		0	0	0
Alphonse Grisel	M	FRA	1896	Artistic Gymnastics		0	0	0
Launceston Elliot	M	GBR	1896	Artistic Gymnastics		0	0	0
Alfred Flatow	M	GER	1896	Artistic Gymnastics		0	0	0
Alfred Flatow	M	GER	1896	Artistic Gymnastics	Gold	1	1	3
Alfred Flatow	M	GER	1896	Artistic Gymnastics	Gold	1	1	3
Alfred Flatow	M	GER	1896	Artistic Gymnastics	Silver	1	0	2
Alfred Flatow	M	GER	1896	Artistic Gymnastics	Gold	1	1	3
Alfred Flatow	M	GER	1896	Artistic Gymnastics		0	0	0
Alfred Flatow	M	GER	1896	Artistic Gymnastics		0	0	0
Carl Schuhmann	M	GER	1896	Artistic Gymnastics	Gold	1	1	3
Carl Schuhmann	M	GER	1896	Artistic Gymnastics		0	0	0
Carl Schuhmann	M	GER	1896	Artistic Gymnastics	Gold	1	1	3
Carl Schuhmann	M	GER	1896	Artistic Gymnastics		0	0	0
Carl Schuhmann	M	GER	1896	Artistic Gymnastics	Gold	1	1	3

Medal Distribution (Cleaned - Gymnastics Only)

Table 9: Medal Distribution After Cleaning (Gymnastics Only)

Medal Value	Count	Percentage (%)
Gold	863	3.02
Silver	818	2.86
Bronze	792	2.77
No Medal	26,081	91.34

Gymnastics Type Distribution

Table 10: Distribution by Gymnastics Type

Gymnastics Type	Number of Records
Artistic Gymnastics	27,768
Rhythmic Gymnastics	754
Trampoline Gymnastics	32

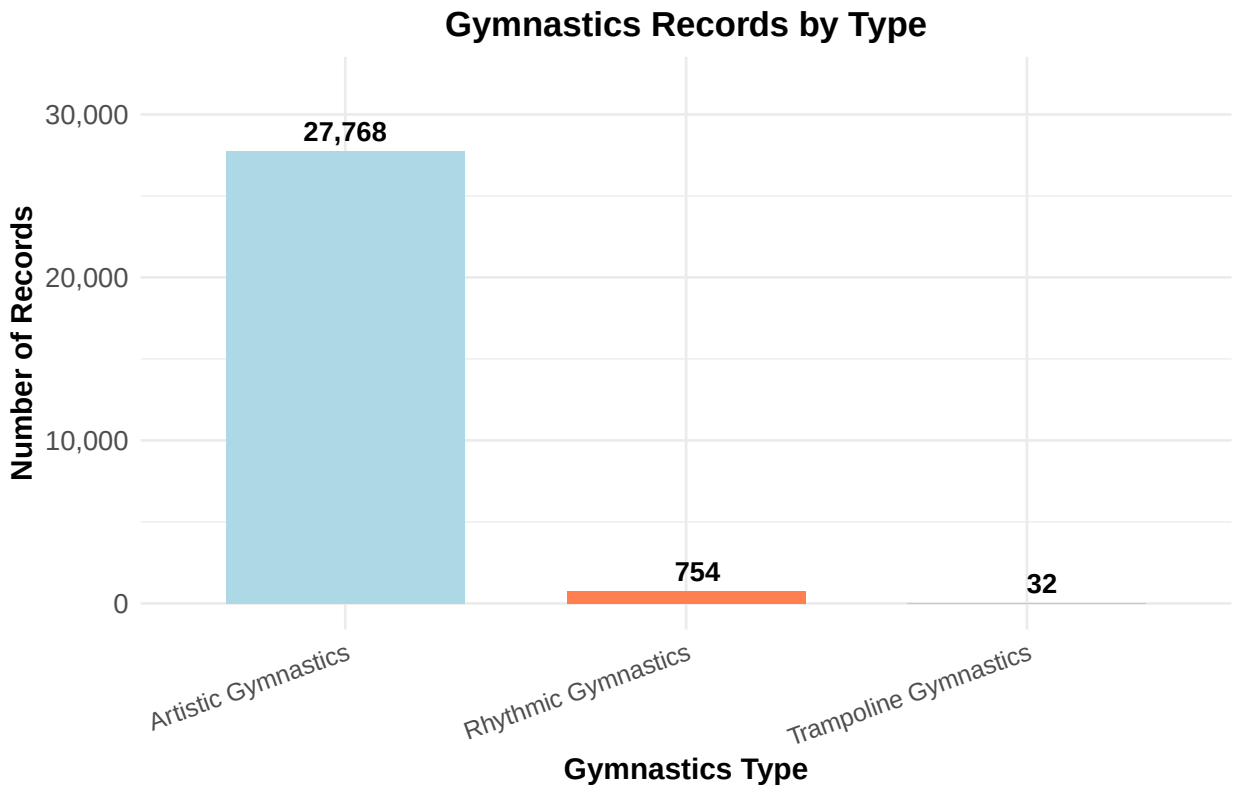

Figure 2: Distribution of Gymnastics Types

New Variables Summary

Table 11: New Medal Outcome Variables

Variable Name	Type	Meaning	Summary
medal_won	Binary (0/1)	1 = won any medal, 0 = no medal	2473 medals
gold_medal	Binary (0/1)	1 = won gold, 0 = did not win gold	863 golds
medal_points	Numeric (0-3)	Gold=3, Silver=2, Bronze=1, None=0	Mean = 0.176

Summary Statistics

Final Gymnastics Dataset Characteristics:

- **Total records**: 28,554 gymnastics performances
- **Unique athletes**: 5,022 individual gymnasts
- **Countries**: 102 different NOCs

- **Years covered**: 1896 - 2020 (30 Olympics)
- **Medal rate**: 8.66% of performances resulted in a medal

Gender Distribution:

Sex	Count	Percentage (%)
F	10,392	36.4
M	18,162	63.6

Key Improvements:

1. Filtered from all-sport records to gymnastics-only dataset
2. Created three quantifiable outcome variables for statistical analysis
3. Preserved original Medal column for verification
4. Standardized gymnastics types into categories
5. Ready for Questions 1 & 2 analysis (home advantage in gymnastics)

Files Created:

- `cleaned_data/athletes_cleaned.csv` - Gymnastics athletes with medal outcome variables

This cleaned variable enables us to answer **Questions 1 & 2**: *"Does hosting the Olympics provide a measurable competitive advantage in gymnastics?"* and *"Does the home advantage differ by gender?"*

Variable 3: GDP Data Reshaping and Per Capita Calculation

1. Before Cleaning

Problem Description

Our research Question 4 investigates whether economic prosperity affects home advantage in gymnastics. To answer this, we need GDP data matched to Olympic years. However, the raw GDP dataset presents major structural and data quality challenges:

Key Issues Identified:

1. **Wide format structure**: GDP data has years as columns (65+ columns from 1960-2024), making it impossible to merge with other datasets
2. **Metadata rows**: First 4 rows contain metadata, not data
3. **Country name inconsistencies**: Same country naming issues as other datasets (no NOC codes)
4. **Missing values**: Many cells are empty strings rather than proper NA values
5. **No per capita calculation**: Only total GDP is provided, but population size varies dramatically across countries
6. **Non-Olympic years**: Includes all years, but we only need Olympic years

Raw Data Sample (Wide Format)

Below is a sample showing the problematic wide format where years are columns:

Table 13: Raw GDP Data - Wide Format (Years as Columns, GDP in Billions USD)

Country	Code	1996	2000	2004	2008	2012	2016	2020
Australia	AUS	401.3	416.2	614.7	1,056.1	1,547.5	1,206.8	1,328.4
Brazil	BRA	850.4	655.4	669.3	1,695.9	2,465.2	1,795.7	1,476.1
China	CHN	868.5	1,223.8	1,984.2	4,667.3	8,673.7	11,456.0	14,996.4
United States	USA	8,073.1	10,251.0	12,217.2	14,769.9	16,254.0	18,804.9	21,354.1

Data Structure Issues

Table 14: Summary of GDP Data Structural Problems

Issue	Problem Description	Impact on Analysis
Data Format	Wide format (years as columns)	Cannot merge with other datasets
Number of Year Columns	65 year columns	Impossible to filter by year
Metadata Rows	First 4 rows are metadata	Headers polluted with non-data
Countries in Dataset	266 country rows	Too many non-Olympic countries
Missing Values	Empty strings instead of NA	Numeric operations fail
NOC Codes	No standardized NOC codes	Cannot join with athletes/hosts data

Summary of Issues:

- **Wide format** with 65 year columns
- Years span 1960-2024, but Olympics only held every 4 years
- Missing values represented as empty strings ""
- No population data for per capita calculation
- Cannot merge with athlete/host datasets in current format

2. Cleaning Process

What We Cleaned

We performed seven major transformations:

1. **Skipped metadata rows**
   - Skipped first 4 rows containing data source information
   - Read from row 5 where actual headers begin
2. **Reshaped from wide to long format**
   - Converted 65+ year columns into rows
   - Created `year` column from column names (X1960 → 1960)
   - Transformed from ~270 rows × 69 columns to ~17,000+ rows × 4 columns
   - Each country-year combination now gets its own row
3. **Cleaned GDP values**
   - Converted empty strings `""` to proper `NA` values
   - Converted GDP from character to numeric type
   - Enables statistical calculations
4. **Filtered to Olympic years only**
   - Matched with summer Olympic years (1896-2020)
   - Removed non-Olympic years
   - Reduces dataset size and focuses on relevant years
5. **Added NOC standardization**
   - Used country codes as NOC identifiers
   - Enables merging with other cleaned datasets
6. **Merged with population data**
   - Loaded population-by-age-group dataset
   - Calculated total population (sum of all age groups)
   - Merged by NOC code and year
7. **Calculated GDP per capita**
   - Formula: `gdp_per_capita = GDP / total_population`
   - Accounts for country size differences
   - Enables fair economic comparisons

Why We Cleaned It

Research Necessity: - **Question 4** asks: "Does economic prosperity (GDP, demographics) affect the home advantage magnitude?" - Need to compare rich vs poor host nations - Raw GDP totals are misleading (China vs Singapore - both wealthy but different scales) - GDP per capita provides meaningful economic prosperity measure

Statistical Requirements: - Cannot merge wide-format data with long-format athlete records - Need year as a variable for time-series analysis - Must standardize across vastly different country populations - Missing value handling required for statistical models

Data Integration: - Reshaping to long format enables joining with athletes and hosts data - NOC codes link all datasets together - Olympic-year filtering ensures data alignment

Expected Outcome

After cleaning, we expect: - Long format with one row per country-year combination - Only Olympic years included - GDP and GDP per capita both available - NOC codes for merging with other datasets - Approximately 3,000-5,000 country-year records - Ready for Question 4 economic analysis

3. After Cleaning

Cleaned Data Sample (Long Format)

Table 15: Cleaned GDP Data - Long Format with Per Capita

NOC	Country	Year	GDP (USD)	Population	GDP per Capita (USD)
AUS	Australia	1,996	401,341,880,621	18,304,158	21,926.2684
AUS	Australia	2,000	416,167,815,093	19,131,060	21,753.5158
AUS	Australia	2,004	614,659,980,083	20,044,992	30,664.0172
AUS	Australia	2,008	1,056,112,427,190	21,368,386	49,424.0617
AUS	Australia	2,012	1,547,532,281,116	22,849,865	67,726.1017
AUS	Australia	2,016	1,206,836,962,282	24,326,444	49,610.0853
AUS	Australia	2,020	1,328,414,058,378	25,739,213	51,610.5158
BRA	Brazil	1,996	850,426,432,992	164,201,321	5,179.1693
BRA	Brazil	2,000	655,448,231,984	174,016,653	3,766.5834
BRA	Brazil	2,004	669,289,424,806	182,672,961	3,663.8670
BRA	Brazil	2,008	1,695,855,083,498	190,364,291	8,908.4727
BRA	Brazil	2,012	2,465,227,802,807	196,871,725	12,522.0003
BRA	Brazil	2,016	1,795,693,482,853	203,211,581	8,836.5706
BRA	Brazil	2,020	1,476,107,231,310	208,652,347	7,074.4818
CHN	China	1,996	868,523,936,530	1,230,942,446	705.5764
CHN	China	2,000	1,223,754,919,971	1,269,576,594	963.9079
CHN	China	2,004	1,984,196,551,300	1,302,093,098	1,523.8515
CHN	China	2,008	4,667,346,414,522	1,333,811,363	3,499.2553
CHN	China	2,012	8,673,664,713,189	1,369,506,753	6,333.4224
CHN	China	2,016	11,456,024,084,962	1,404,029,400	8,159.3905
CHN	China	2,020	14,996,414,166,715	1,426,071,437	10,515.8927
USA	United States	1,996	8,073,122,000,000	270,817,511	29,810.1920
USA	United States	2,000	10,250,952,000,000	281,440,256	36,423.1903
USA	United States	2,004	12,217,196,000,000	292,740,259	41,733.9113
USA	United States	2,008	14,769,862,000,000	304,917,598	48,438.8638
USA	United States	2,012	16,253,970,000,000	317,055,101	51,265.4423
USA	United States	2,016	18,804,913,000,000	329,119,530	57,137.0316
USA	United States	2,020	21,354,105,000,000	339,375,643	62,921.7371

Before vs After Transformation

Table 16: GDP Data Transformation Summary

Aspect	Before Cleaning	After Cleaning
Data Format	Wide (years as columns)	Long (one row per country-year)
Number of Rows	271 countries	4,522 country-year combinations
Number of Columns	69 columns	6 columns
Years Included	All years (1960-2024)	Olympic years only (1960-2024)
GDP Variable Type	Character (with empty strings)	Numeric
Missing Value Format	Empty strings “ ”	Proper NA values
Population Data	No population	Total population included
GDP per Capita	Not available	Calculated (GDP / population)
NOC Standardization	No NOC codes	NOC codes added
Ready for Merge	No - incompatible format	Yes - can merge by NOC + year

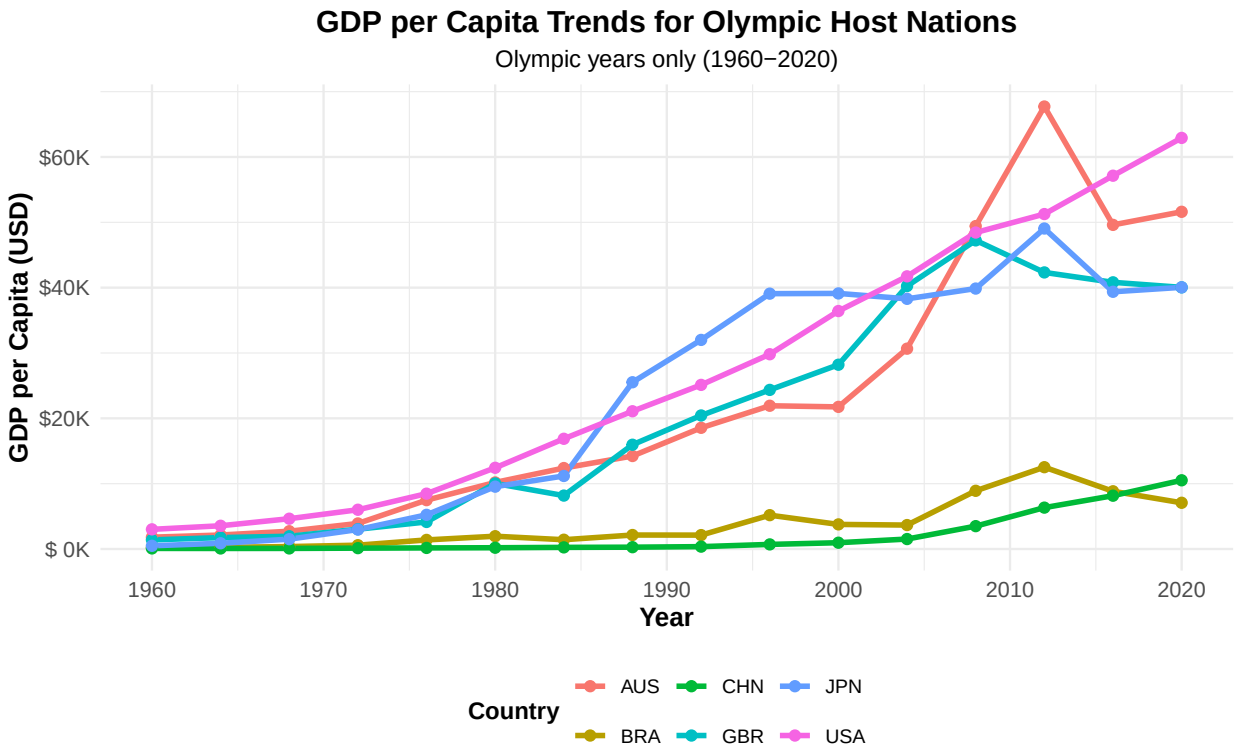

Figure 3: GDP per Capita Trends for Some Selected Olympic Host Nations

GDP per Capita Comparison (2016 Olympics)

Table 17: GDP per Capita in 2016 (Rio Olympics) - Top 10 + Brazil

NOC	Country	GDP per Capita (USD)
MCO	Monaco	175,089.037
LIE	Liechtenstein	165,749.038
BMU	Bermuda	108,726.793
LUX	Luxembourg	106,636.191
CHE	Switzerland	82,151.505
IMN	Isle of Man	82,056.422
CYM	Cayman Islands	77,777.602
MAC	Macao SAR, China	71,059.902
NOR	Norway	70,859.686
IRL	Ireland	64,322.621
BRA	Brazil	8,836.571

Summary Statistics

Final GDP Dataset Characteristics:

- **Total records**: 4,522 country-year combinations
- **Countries**: 266 unique NOCs
- **Years covered**: 1960 - 2024
- **Olympic years**: 17 Olympic years

- **Records with GDP**: 3,766 (83.3%)
- **Records with GDP per capita**: 2,793 (61.8%)

Data Coverage by Era:

Table 18: GDP Data Coverage by Time Period

Time Period	Has GDP Data	Has GDP per Capita
1960-1979	861 (64.7%)	649 (48.8%)
1980-1999	1131 (85%)	898 (67.5%)
2000-2024	1774 (95.3%)	1246 (66.9%)

Key Improvements:

1. **Structural transformation**: Wide format (unusable) → Long format (mergeable)
2. **Data reduction**: From 65+ year columns to focused Olympic years only
3. **Value creation**: Added GDP per capita variable for meaningful comparisons
4. **Integration ready**: NOC codes enable joining with athletes and hosts datasets
5. **Statistical readiness**: Numeric types, proper NA handling, time-series ready

Enables Question 4 Analysis:

This cleaned variable allows us to investigate: *"Do wealthier host nations experience different home advantage effects? Does the economic prosperity of a country affect their gymnastics performance when hosting?"*

We can now compare: - High vs low GDP per capita host nations - GDP trends over time for repeat hosts - Economic prosperity's correlation with home advantage magnitude

Files Created:

- `cleaned_data/gdp_long.csv` - Long format GDP with per capita calculations