Data Cleaning Report

Olympics Gymnastics Home Advantage Analysis - Team 5

Han Nguyen, Matthew Bazzell, Carlos Flores

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

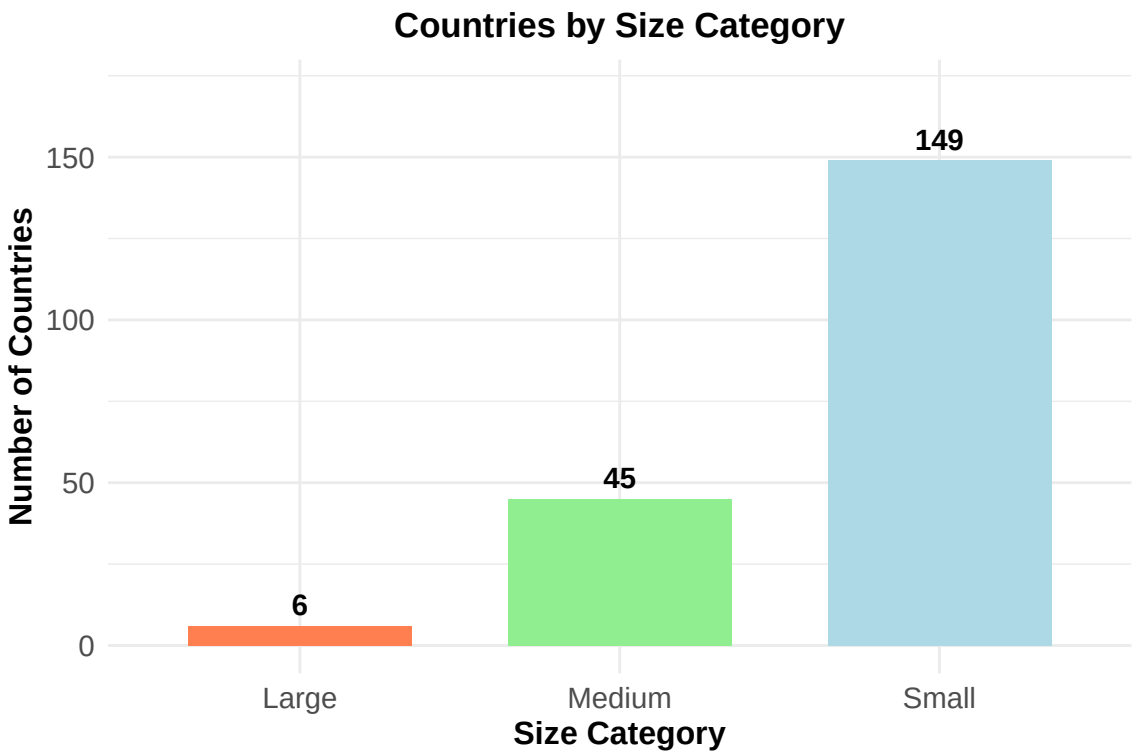

Figure 1: Distribution of Country Size Categories

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

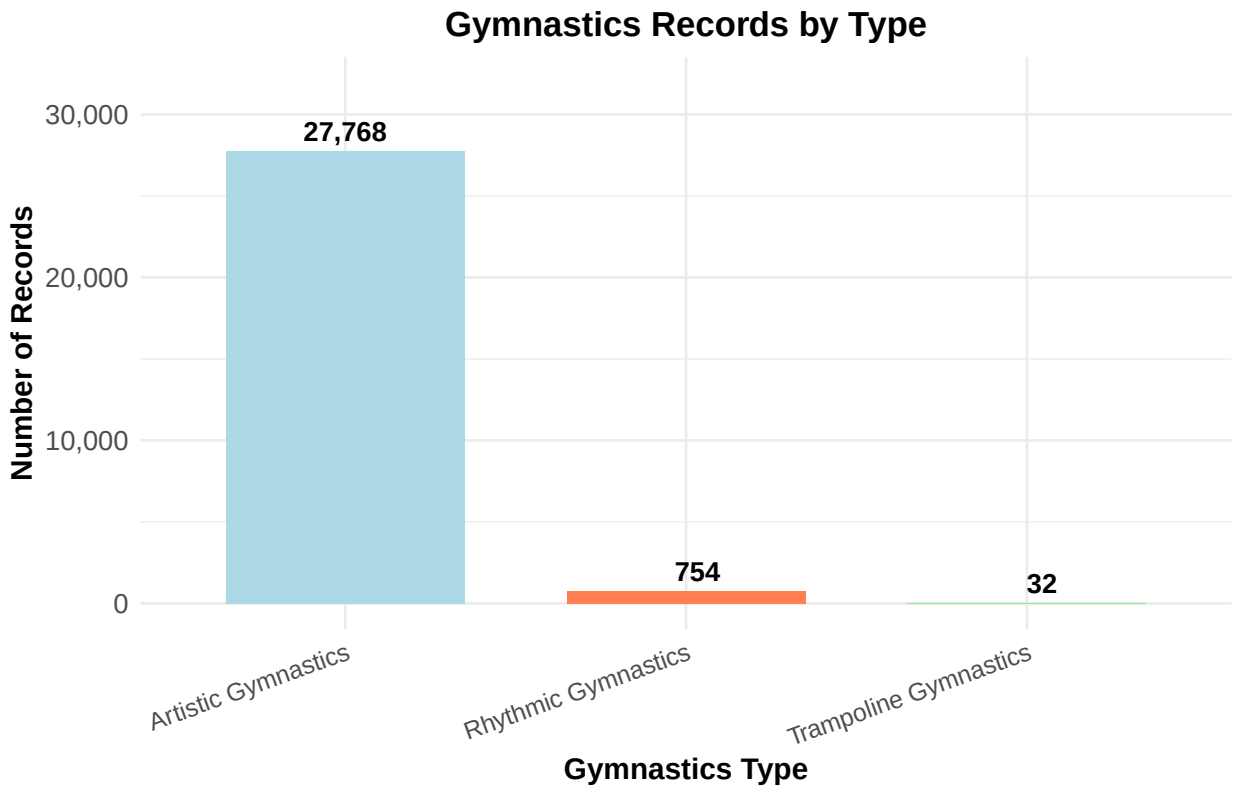

Figure 2: Distribution of Gymnastics Types

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?"*