ORIE 3120 Group 5: Milestone 2

Skylar Weirens, David Valarezo, Anna Geng, Carina Lau

The Olympics represent the peak of global athletic competition, but they also raise important questions about fairness, opportunity, and international parity. Our group began by exploring a comprehensive dataset covering over a century of Olympic history from 1896 to 2016 with more than 250,000 athlete entries. Each row reflects a single athlete's participation in a specific Olympic event, including key details like age, country, sport, and medal outcome.

As we explored the data, we became increasingly interested in the question of fairness both in terms of geopolitical influence and the structure of Olympic competition itself. We used linear regression and other statistical tools to investigate patterns and possible biases. Key questions that guided our analysis included:

1. Are countries winning a fair number of gold medals relative to their total medal count?
2. Is judging biased? Are some nations systematically advantaged in subjective sports like gymnastics compared to objective events like track or swimming?
3. Are the Summer and Winter Games equally fair? Do certain countries dominate one season more than the other, and what does this reveal about Olympic funding distribution?
4. Which events are the most inclusive? We identified the sports with the most diverse set of gold medal-winning countries to understand where international access and opportunity are highest.

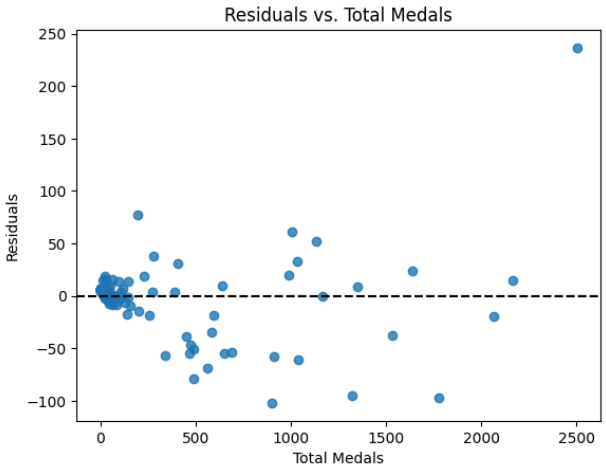
These questions helped us move beyond medal counts and into a deeper analysis of what fairness looks like at the Olympics. Before drawing conclusions, we first check and validate the assumptions of linear regression to ensure our statistical analysis is sound when necessary.

Check Assumptions for Linear Regression

The validity of our analyses depends on whether the following assumptions are satisfied. More specifically, the residuals must:

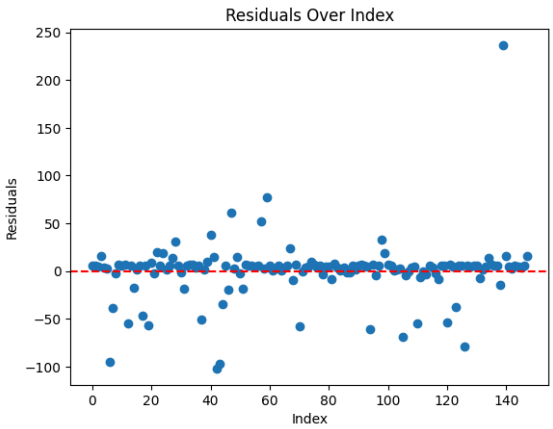
Have Constant Variance

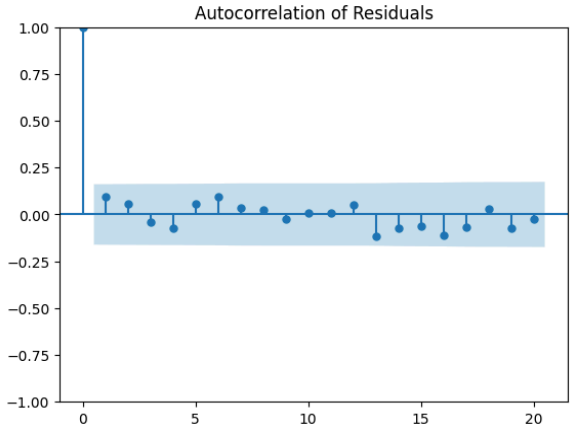
To assess the assumption of constant variance, we plotted residuals against the predicted values of gold medals. The spread of residuals remains relatively consistent across the range of predicted values. While there are a few outliers, particularly at higher predicted values, the bulk of the residuals are symmetrically distributed around zero. This suggests that the variance of the errors does not depend on the size of the prediction, supporting the assumption of homoscedasticity. Given that larger countries naturally have greater variability in medal counts, the mild increase in spread is reasonable and expected in this context.



Be Independent

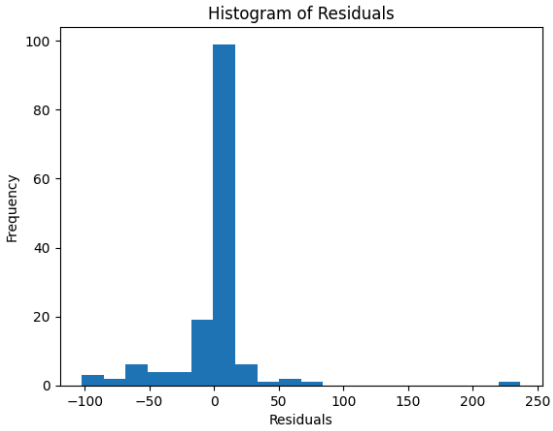
We checked for residual independence using two methods. Firstly, we plotted the residuals over the index which showed no clear trend or pattern, which suggests randomly distributed residuals. Secondly, we made an autocorrelation plot that showed that all the lag coefficients fell within the 95% confidence interval. This shows that the residuals are not correlated, thus supporting the assumption of mutual independence.

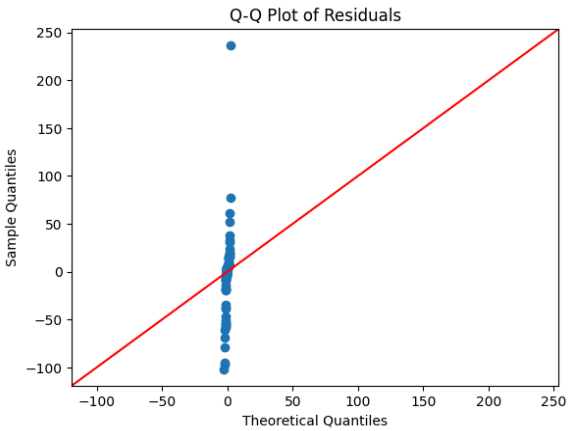




Be Normally Distributed

The histogram on the bottom left shows a sharp peak at approximately zero and a longer right tail, which demonstrates that most residuals are small, with a few large errors that skew the distribution. Moreover, the Q-Q plot further confirms this by showing deviation from the red diagonal line on both tails. While the residuals appear heavy-tailed (not perfectly normal), their overall symmetry and centering around zero suggest reasonable model stability. Therefore, we proceed with the analysis, acknowledging this minor deviation.

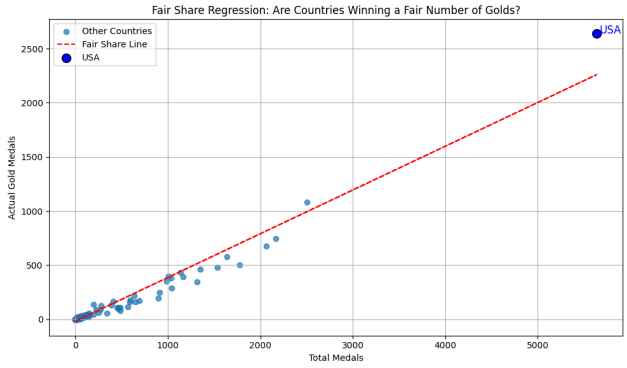


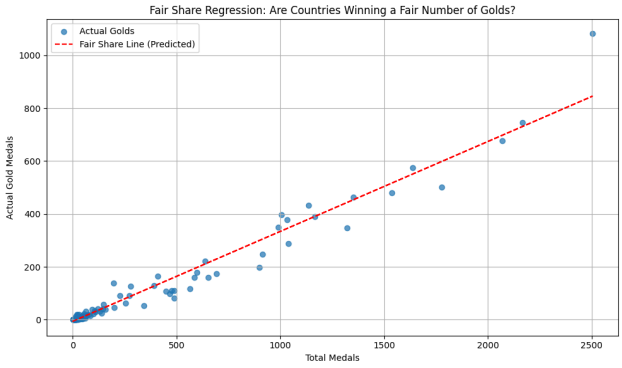


Linear Regression: Fair Share of Gold Medals

In the context of fairness in the Olympics, we first examined whether or not we can predict the number of gold medals a country wins based on their total medals. In other words, we will explore through linear regression, whether countries are winning a fair number of golds in comparison to one another. We created a linear regression model that predicts the number of gold medals a country should win based solely on their total number of medals. In both graphs, the red dashed line represents this expected "fair share" of golds — countries above the line are overperforming, while those below are underperforming relative to the global trend.

In the left graph, we highlight the United States, which won significantly more golds than the model predicts (greatly above the red line), while also being an extreme outlier in the total medal count, raising questions about whether certain countries have structural advantages in converting medal opportunities into golds. This method allows us to visualize performance gaps not just in terms of totals, but in fairness: given a country's overall success, are they being rewarded with gold medals at the same rate as others?





R²: 0.959
Slope: 0.404
Intercept: -18.130

R²: 0.962
Slope: 0.340
Intercept: -5.936

Figure 1: This figure presents a linear regression predicting the number of gold medals a country should win based on their total medal count. The red line represents the expected trend. The United States is highlighted in blue on the left and removed from the model on the right; note the differences in scale.

The right graph visualizes the same concept but removes the United States as an outlier to better examine how the rest of the countries align with the fair share regression trend. By doing so, we reduce the distortion caused by the United States' extremely high medal count, which can skew the scale and make it more difficult to see meaningful differences among the other countries. Moreover, the R^2 value increases slightly when the United States is excluded (from ~0.959 to ~0.962), showing that while the model still fits well, the United States' extreme overperformance introduces more variation around the trend. Removing it provides a clearer view of how fairly golds are distributed among the rest of the countries. It reveals that most countries follow a fairly linear relationship between total medals and gold medals, suggesting a relatively consistent gold conversion trend. However, some countries still fall noticeably above or below the line, indicating that even among non-outliers, disparities in gold medal outcomes exist.

Overall, these disparities raise broader questions about equity in international competition — including how access to resources, funding, training facilities, and historical dominance may influence performative outcomes in the Olympics. This is particularly evident in the case of the United States, which appears as a clear outlier in the first graph, significantly overperforming in gold medal conversion relative to the global trend.

Linear Regression and Judging Bias

Sports such as swimming, weightlifting, and athletics have well defined rules for who wins. The fastest runner, the strongest weightlifter, or the fastest swimmer takes home the gold with little exceptions ever being made. However, these standards are not as clear when it comes to gymnastics. As an inherently subjective sport, despite decades of attempts to eliminate bias, cheating through biased judging has been a source of contention throughout the sport's history.

In order to explore the possibility of biased judging, we looked at a number of different countries who were significant performers (who won the most medals) in Gymnastics over the years. We then analyzed their performance in gymnastics compared to other sports. If there was a consistent overperformance in gymnastics, this opens the possibility that there is biased judging towards (or against) a country. When selecting our countries and other sports we wanted to reduce the amount of possible external reasons for difference in performance outside of biased judging. For instance, we selected swimming, athletics, and weightlifting because they are entirely objective, have a multitude of events per Olympics, and most importantly all have low barriers of entry. It makes less sense to compare performance to sports such as hockey which has few events to win and has high barriers of entry (high cost of equipment and facilities). Lastly, while it is important that there are many events in our chosen sports to increase sample size, they all have varying numbers of events. To combat this we weighted the number of events to be equal. This means that despite there being more events in athletics, the weight per medal is scaled equally with the amount

of events in gymnastics. This approach ensured that medal counts reflect comparable opportunity, not simply volume of competition.

When analyzing the data concerning the United States and China, the results revealed largely no statistically significant difference between judged and non-judged events. This means that the chance of unfair judging giving these countries an advantage is extremely low. However, this is not the case when looking at the Soviet Union's performance during its existence. By inspection, graphs show a clear difference in performance in gymnastics when compared to athletics, swimming and weightlifting. There is not a single instance of overlap between gymnastics and other sport performance lines as seen in (*Figure 2.1-2.3*).

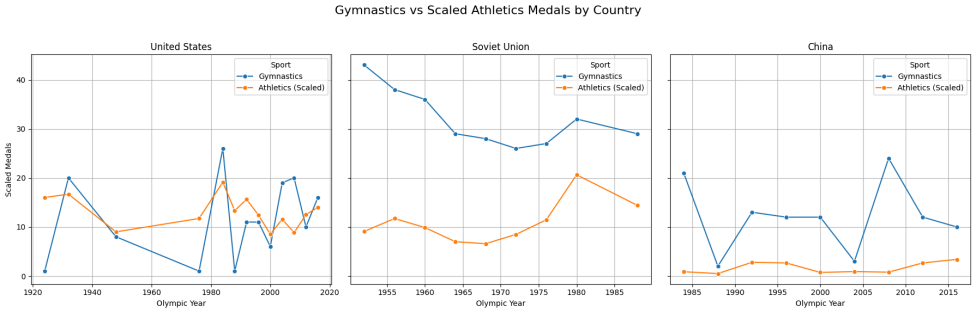


Figure 2.1: This figure compares the number of scaled medals won in gymnastics and athletics for the United States, Soviet Union, and China during the Olympics.

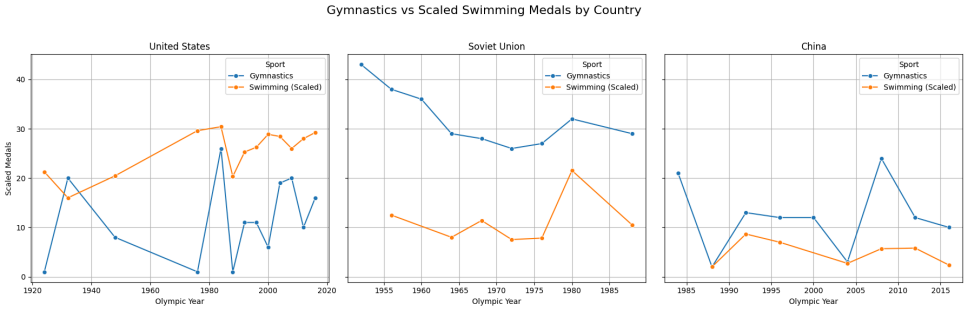


Figure 2.2: This figure compares the number of scaled medals won in gymnastics and swimming for the United States, Soviet Union, and China during the Olympics.

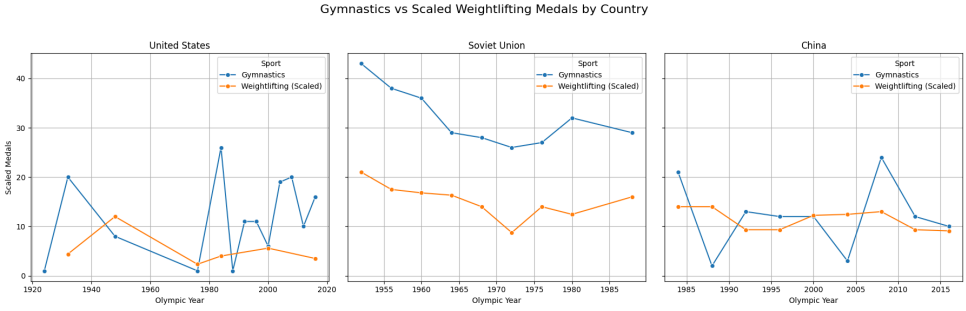


Figure 2.3: This figure compares the number of scaled medals won in gymnastics and weightlifting for the United States, Soviet Union, and China during the Olympics.

Ordinary linear regression also aligns with our analysis by inspection. Looking at the table of resulting P-values and R^2 values also points to evidence of possible biased judging for the Soviet Union. To make the graph more legible, a "Judging Bias" column was created where green symbolizes no bias and red symbolizes possible

bias. The Soviet Union consistently outputs low P-Values and high R^2 values as seen in (*Figure 3*) meaning there is a possibility of biased judging in their favor.

Comparison	Country	Judged Coef	P-Value	R^2	Significance	
Gymnastics vs. Athletics	USA	-1.489	0.143	0.015	Insignificant P and R^2	
Gymnastics vs. Athletics	Soviet Union	18.37	0	0.845	Significant P and R^2	
Gymnastics vs. Athletics	China	10.41	0.001	0.536	Significant P Insignificant R^2	
Gymnastics vs. Weightlifting	USA	6.23	0.095	0.155	Insignificant P and R^2	
Gymnastics vs. Weightlifting	Soviet Union	16.8	0.001	0.777	Significant P and R^2	
Gymnastics vs. Weightlifting	China	0.69	0.786	0.005	Insignificant P and R^2	
Gymnastics vs. Swimming	USA	-13.85	0	0.543	Significant P Insignificant R^2	
Gymnastics vs. Swimming	Soviet Union	20.668	0	0.803	Significant P Significant R^2	
Gymnastics vs. Swimming	China	7.22	0.024	0.313	Significant P Insignificant R^2	

Figure 3: This table presents the linear regression coefficients, p-values, and R^2 values for gymnastics compared to athletics, weightlifting, and swimming across three countries. Red represents the possibility of judging bias and green represents no possibility of bias.

An interesting aside: If you look closely at the United State's performance in Swimming over the years you may notice a significant P-value, but a negative judged coefficient. This result would typically signal biased judging, but swimming is an entirely objective sport. The trend instead most likely signals the domination in swimming by the United States with athletes such as Michael Phelps who would have contributed to the overperformance of the United states in the early 2000's.

Why couldn't this aside be applied to the Soviet Union's overperformance?

When applying the same linear regression to the entire dataset, there were almost no countries where potential bias was identified. However, one of the few that were identified as having potential bias was East Germany. This is significant because despite being an independent nation, East Germany was immensely influenced by the Soviet Union[1]. Although it competed under its own flag, the political and ideological alignment between East Germany and the USSR was extremely tight, particularly during the Cold War. This alignment extended beyond diplomatic affairs and into realms like sports, where showcasing national superiority on the world stage was a priority. As a result, the same systems of influence and pressure that may have affected judging for Soviet athletes could plausibly extend to East German athletes as well. A weaker, but similar, trend to the USSR is seen in (*Figure 4*) as well as statistical significance (*Figure 5*) is reflected in East Germany's performance. This overperformance by East Germany not only supports the claim of biased judging through statistical analysis, but also geopolitical relationships during the time period.

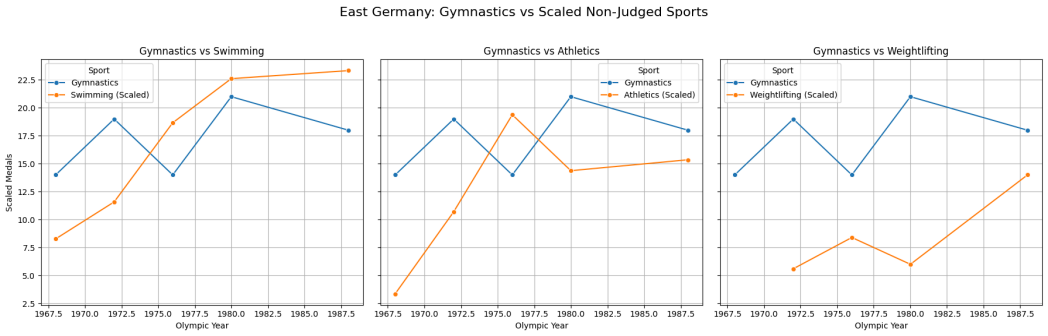


Figure 4: This figure compares the number of scaled medals won in gymnastics compared to swimming, athletics, and weightlifting for East Germany in the Olympics.

[1] https://history.state.gov/countries/german-democratic-republic

Sport Comparison	Judged Coef	P-Value	R^2
Gymnastics vs. Swimming	0.92	0.004	0.602
Gymnastics vs. Athletics	1.15	0.002	0.685
Gymnastics vs. Weightlifting	0.78	0.01	0.541

Figure 5: This table presents the linear regression coefficients, p-values, and R^2 values for East Germany compared to athletics, weightlifting, and swimming. These values are a weaker, yet still statistically significant comparison to the Soviet Union's performance.

Seasonal Dominance and Fairness

By examining countries' medal shares across both Summer and Winter Games, we aim to uncover their natural "peer groups": nations that dominate both seasons, those that specialize in either Summer or Winter and those that exhibit a balanced approach. Understanding these groups allows us to celebrate diverse achievements while respecting differences in geographic, climatic, and economic contexts, ensuring fairness in how excellence is recognized and rewarded.

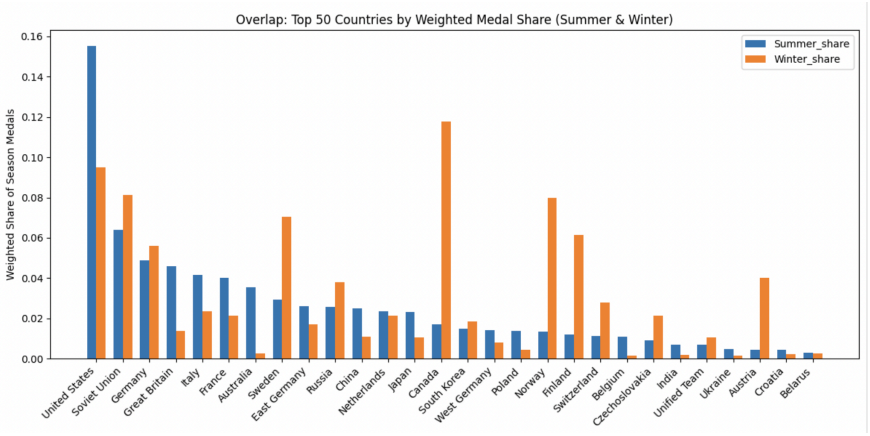


Figure 6: This visualization shows the 28 countries that rank in the top 50 by weighted medal share in both the Summer and Winter Olympics, based on a medal scoring system that weights gold as 3 points, silver as 2, and bronze as 1. Each country is represented with two bars: one for its share of total Summer medals and one for Winter.

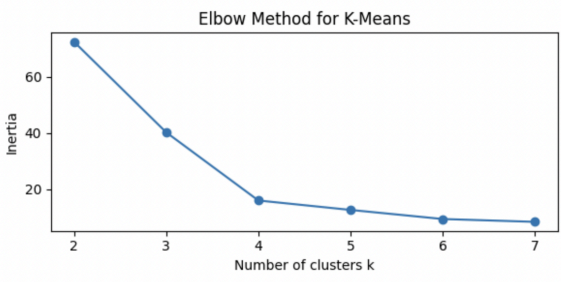
The countries shown in *Figure 5* clearly demonstrate dominance in both seasons, but with varying degrees of strength. Some, like the United States, Great Britain, and Australia, have much stronger performances in the Summer Games. Others, such as Canada, Finland, and Austria, show the opposite pattern, with their weighted medal shares heavily concentrated in the Winter Olympics. A few countries like the Soviet Union, Germany, and Switzerland exhibit a more balanced distribution. This suggests that while these countries are elite performers year-round, their seasonal strengths differ, reflecting factors like climate, investment priorities, and historical focus in specific sports.

To continue our analysis, we start by aggregating each nation's medals into a unified medal-point score (3×Gold + 2×Silver + 1×Bronze) separately for Summer and Winter, and then normalized those scores by dividing by the global total medal points in each season to produce Summer_share and Winter_share. By restricting to countries with at least one point in both seasons, we ensured every plotted point was meaningful.

Instead of applying a supervised regression, we chose K-means clustering, an unsupervised algorithm that discovers latent groups without requiring a predefined outcome. K-means works by initializing k centroids at random in feature space, assigning each country to its nearest centroid (using Euclidean distance), then recomputing each centroid as the mean of its assigned points. We iterate these steps until cluster membership stabilizes, thereby minimizing within-cluster variance. Because Winter_share and Summer_share can have different variances, we first

applied z-scaling. We subtracted each feature mean across all countries and divided it by its standard deviation, ensuring both dimensions contribute equally to distance calculations. To choose the number of clusters k, we began by standardizing our two features so that each has mean zero and unit variance, yielding an n×2 data matrix X. Then, for each candidate cluster count k=2,3,…,7, we ran K-means on X and recorded its within-cluster sum of squares (often called inertia) using the formula below:

$$W(k) = \sum_{i=1}^{k} \sum_{x \in Ci} \|x - \mu i\|^2$$



where Ci is the set of points assigned to cluster i and μi is that cluster's centroid. We plotted W(k) against k and observed a sharp decline from k=2 through k=4, after which the curve flattens: additional clusters beyond four produce only marginal reductions in inertia. This "elbow"— the point at which the rate of decrease in W(k) markedly slows — appears clearly at k=4. Because adding more clusters past that point yields diminishing returns, we chose 4 as the optimal number of clusters.

The resulting four clusters show clear archetypes. The "Dominates Both" group holds the few nations in the top 10% of both Summer_share and Winter_share, representing actual all-season superpowers like the United States and Soviet Union. Summer-dominant and Winter-dominant clusters capture countries whose share profiles exceed the other season by more than 0.5 percentage points such as Canada with its winter domination or Great Britain with its summer domination. Lastly, the Average cluster contains those with roughly balanced shares which shows average overall performance. We labeled the top five medal-point countries in each category to present visualization with concrete examples. We also overlaid threshold lines to make category boundaries transparent. We show the 45° balance line that illustrates the perfect balance of a country earning the same proportion of medals in both seasons, and add the 90th-percentile bars and ±0.5 pp bands around the 45° balance line to show area for different clusters.

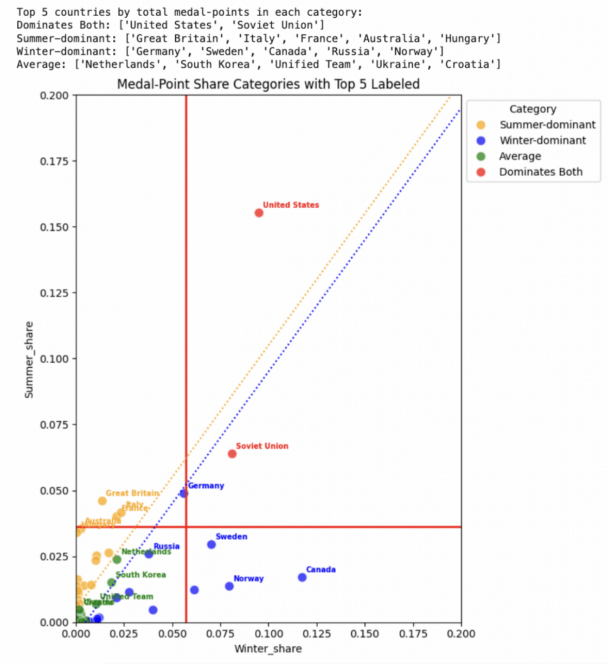


Figure 7: This graph shows which countries dominate in which seasonal sports are divided into four quadrants.

Unlike regression techniques studied in class, which require specifying a target variable and fitting a single predictive relationship, K-means uncovers natural performance archetypes, enabling organizations to benchmark countries against true peers, allocate resources more equitably and craft context-aware narratives of Olympic success.

This clustering supports our broader fairness analysis and highlights possible structural advantages such as geography. However, the clustering also points out how totally dominant nations most likely aren't dominant because of structural advantages, but investments. This is important when looking at fairness because it shows how Olympic outcomes may not be the result of talent, but instead where money is spent. Extensive sports budgets may then undermine the potential of athletes or nations due to their lack of funding.

Most Fair Sports

Given the vast number of participants and events in the Olympics, we wanted to know which events have experienced the greatest number of gold medalists from distinct countries.This is important for several reasons. For athletes, it pinpoints events where the path to gold is genuinely feasible rather than barred by infrastructure or historical dominance. For fans it highlights the most unpredictable and exciting events. Lastly, this information could lead to initiatives to nurture talent from underrepresented regions, strengthening the competitive environment.

To find the most "open" events, we made a bar chart of the 20 events with the highest number of distinct gold medaling countries throughout the last 120 years. First we filtered the dataset by only keeping the records

where an individual won a gold medal. We used the "Event" field to group our data and used the "NOC" field to count the number of distinct countries that have won gold for each event. We plotted the top 20 events in *Figure 8.*

We also wanted to see how this compares to a truly fair competition where every participating country has an equal probability of winning. To do this, we ran a Monte Carlo simulation where we randomly selected a participating country to win the gold medal with uniform probability for each edition of an event. We summed the number of distinct countries that won gold for each simulation and took the average of this sum over all 5,000 simulations. We performed this process for each of the 20 events we found earlier. The result is the red vertical line which represents the expected number of distinct countries that won gold for each of the top 20 events (*Figure 8)*.

Upon first glance, we can see that the expected result is higher than the actual result for 9/20 events. This means that these events aren't as "fair" as our simulation and that more countries would win gold if each participating country had an equal chance of winning. Drawing gold medal winners from a uniform distribution produced anywhere from 1 to 4 more distinct countries that won gold. Clearly, for these events there were countries that won multiple times. The remaining 11 events are actually more "fair" than our simulations since more countries win than if medals were distributed uniformly. For these events, drawing from a uniform produced anywhere from 0 to 6 fewer winners. Also, note that the top 20 events don't have as many barriers to entry such as sports like ice hockey or swimming, which require significant infrastructure to even begin.

To investigate our results further, we defined our Null Hypothesis**:** the gold medals were assigned under a uniform distribution. We performed a one-tailed hypothesis test and using a significance level $\alpha = 5\%$, we estimated the p-values for each event - the probability of the simulated result being as extreme as our observed result. Note that T is the simulated result and T_{Obs} is the actual result from our data.

$$P\text{-}value = (T \geq T_{Obs} | H_0) = \frac{\#\ of\ simulations\ where\ result \geq actual\ result}{Total\ Simulations}$$

The resulting p-values are shown in *Figure 9.* Interestingly, men's football has p-value = 0 meaning that out of our 5,000 simulations, there were close to 0 simulations where the simulated number of distinct countries that won a gold medal (T observed) was greater than the actual number of distinct countries that won a gold medal (T). This means soccer is more diverse than chance, which reflects the sport's global popularity and large participant pools. On the other hand, men's athletics has a lot less distinct winners than expected. The event is fair since there is a low barrier to entry for running. However, powerhouses like Ethiopia, which has won gold 6 times, have contributed to a much lower result than expected.

Given there are so many events in the Olympics, it's interesting to see which ones are the most open or anyone's game. Events that fall below the uniform benchmark reveal where strategic investments like equipment costs, specialized training, or historical dominance limit competitive access. Conversely, events that exceed random expectations demonstrate level playing fields where any nation has a viable path to gold. By quantifying "openness," policymakers and sports federations can identify which competitions may benefit from targeted development programs, ultimately fostering a more equitable Olympic experience.

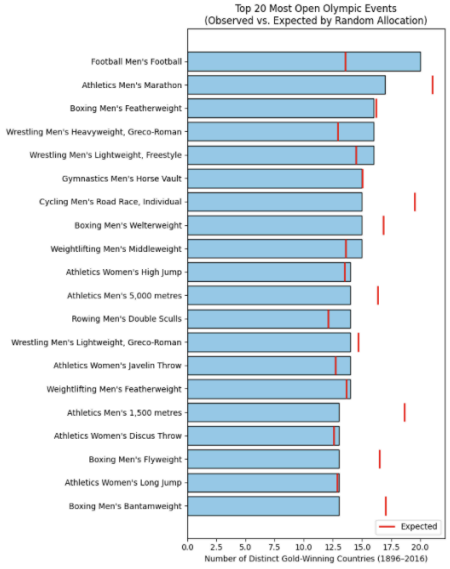


Figure 8: Horizontal bar chart showing the top 20 Olympic events ranked by the total number of distinct gold medal winning countries (1896–2016). Red vertical lines indicate the expected counts.

	Event	Appearances	ObservedDistinct	ExpectedMean	StdSimulated	p_value
0	Football Men's Football	27	20	14	1.379626	0.0000
1	Athletics Men's Marathon	29	17	21	1.841391	0.9930
2	Boxing Men's Featherweight	23	16	16	1.618704	0.6650
3	Wrestling Men's Heavyweight, Greco-Roman	26	16	13	1.402098	0.0290
4	Wrestling Men's Lightweight, Freestyle	24	16	14	1.538388	0.2590
5	Boxing Men's Welterweight	24	15	17	1.644358	0.9202
6	Cycling Men's Road Race, Individual	26	15	20	1.732604	0.9982
7	Gymnastics Men's Horse Vault	24	15	15	1.608368	0.6274
8	Weightlifting Men's Middleweight	23	15	14	1.489093	0.2824
9	Athletics Men's 5,000 metres	24	14	16	1.646404	0.9602
10	Athletics Women's High Jump	21	14	14	1.500460	0.4968
11	Athletics Women's Javelin Throw	20	14	13	1.427454	0.2902
12	Rowing Men's Double Sculls	24	14	12	1.361639	0.1486
13	Weightlifting Men's Featherweight	23	14	14	1.473479	0.5456
14	Wrestling Men's Lightweight, Greco-Roman	25	14	15	1.549013	0.7794
15	Athletics Men's 1,500 metres	29	13	19	1.740767	0.9996
16	Athletics Women's Discus Throw	21	13	13	1.465402	0.5190
17	Athletics Women's Long Jump	18	13	13	1.413829	0.6142
18	Boxing Men's Bantamweight	25	13	17	1.692769	0.9962
19	Boxing Men's Flyweight	24	13	16	1.638670	0.9936

Figure 9: P-values for these 20 events under the null hypothesis of uniform gold medal allocation. A one-sided test estimates how often random draws produce as many or more distinct winners than observed. Values below 0.05 indicate less openness than chance, while values above 0.95 indicate greater openness.