Problem Chosen	2025 MCM/ICM Summary Sheet	Team Control Number
C		2522820

The Best of the Best: Olympic Medal Tables

Summary

Winning at the Olympics is the pinnacle of physical achievement for any athlete. It is an honor to even be invited to the Olympics, and it is a global spectacle for weeks to see the best athletes of the modern age compete to be the best in the world in a large variety of sports. As developing countries rise in the athletic scene, we are seeing a large increase in the number of participants and countries who achieve their first Olympic medal despite multitudes more setbacks than traditionally competing countries.

As modelers, we have been tasked with predicting the outcome of the 2028 Summer Olympics held in Los Angeles, United States. Olympic medal data exhibits censorship in the form of zero inflation, so we first consider a **Tobit Model** that considers **random noise** and **unobserved random effects**, optimized with a **Maximum Likelihood Estimate**. From regression analysis on the Tobit Model, we discovered it suffered from **heteroscedasticity**, which is detrimental to the Tobit model. So, we implement a **Hierarchical Mixed Linear Effects Model** to predict medal counts based on **Host Status**, **Number of Athletes**, and **Past Performance**. Country performance was split into clusters using **Density-Based Spatial Clustering of Applications with Noise (DBSCAN)**, splitting the countries into three clusters. The Mixed Linear Effects Model is then used to perform a **Negative Binomial Regression** based on the predictors, as well as a **Probit Regression** in the low-performing cluster to determine probability of earning 0 medals.

We predicted the Top 3 (in gold count) countries in the 2028 Summer Olympics to be the **United States** (48), **Germany** (28), and **China** (25), each with a 95% confidence interval range of around 30. The top three countries that we predict will improve the most in terms of ranking is **Jamaica** (+27), **Germany** (+8), and **Ukraine** (+8), and the worst performance changes in ranking we predict are **Canada** (-99), the **Netherlands** (-31), and **South Korea** (-10). Finally, our model also predicts that the countries that will earn their **first** Olympic medal in 2028 are Angola and Saar, with odds of 32.54% and 40.08%, respectively.

After analyzing the dynamics of the rankings, we wanted to understand the impact of events and sports on the rankings. To do this, we created a ranking system of Gold 6 points, Silver 3 Points, and Bronze 2 points. Then, through **Singular Value Decomposition, Sport-Point Efficiency**, and **Revealed Comparative Advantage**, we were able to understand the influence sports and events had on the ranking of various countries.

We also analyzed past performances at the Games to identify possible instances of the great coach effect, or the influence of a great coach in leading teams to anomalous successes. To build our model, we noticed that performances at the Games are time series, so we used a time series analysis incorporating a moving average. We found that the 1976 Romanian Handball team, the 2008 Canadian trampolining team, and the 2000 Russian diving team were the most likely to have experienced the great coach effect.

By using a contribution metric, we used the same model to find the country-sport pairs where the great coach effect was likely to have made the greatest impact on the country's performance. We found that the 1992-2012 Swedish handball teams, the 1984-2004 United States synchronized swimming teams, and the 2000-2016 Russian modern pentathlon athletes likely benefited the most from the great coach effect, with the Swedish handball and United States synchronized swimming athletes earning 100% of their medals during segments where they were led by great coaches, according to our model. Similarly, by using a relative contribution metric, we suggested that Russia invest in a great weightlifting coach, China invest in a great trampolining coach, and Hungary invest in a great weightlifting coach. Based on their performances under great coaches, these investments could lead to up to 20 points for Russia, 14 points for China, and 6.8 points for Hungary.

All of these results were cross-referenced with historical Olympic coaches, and many of our findings lined up with the leadership of coaches that may be considered great.

There are a multitude of benefits in this research—such as understanding how to best allocate resources—but more than that, just as the Games represent the human desire for unity in excellence, these models represent our fundamental desire to understand the world.

Contents

1 Introduction

"The Olympics remain the most compelling search for excellence that exists in sport, and maybe in life itself." —Dawn Fraser, eight-time Olympic medalist.

The Olympic tradition traces back almost three millennia to Olympia in Ancient Greece. Even in an era of constant conflict between the city-states, the ancient Games represented humanity's unified pursuit of excellence. The Ancient Greeks even established a temporary Olympic Truce to ensure the safety of the athletes: unity in competition took precedence.[1]

The modern Olympic Games, established in 1896, have inherited this humanistic spirit. The modern Games have gathered the best athletes from around the world in their common pursuit not just of athletic excellence, but of reaching the limits of human potential and bringing our culturally diverse world closer together in the process. The models we present in this paper aim to understand the factors behind these competitions.

1.1 Problem Restatement

1. (a) Predict the medal table for the Los Angeles 2028 Summer Olympics. Provide prediction intervals for all medal predictions, and predict which countries will improve the most and worsen the most based on their 2024 performance.
   (b) Then, predict which countries will earn their first Olympic medal ever in the 2028 Olympics, providing odds for the estimate.
   (c) Consider how different sports and sports specialties play into country performance in a given Olympics and explore how events chosen by hosting countries affect results.
2. (a) Create a mathematical model that finds potential cases of the "great coach effect," or the positive influence of a great coach on a country's performance in a discipline.
   (b) Use this model to estimate the impact of such potential great coaches on their countries' performances.
   (c) Interpret the results of this model to identify three country-sport pairs that great coaches would bring most value to, and estimate the potential impact of hiring great coaches for these three country-sport pairs.
3. (a) Discuss original insights about Olympic medal counts, and explain how these insights can be used to inform Olympic Committee decisions for future Olympics.

2 Data Pre-Processing

To easily analyze the given data, the CSV files datasets given by the **International Olympic Committee (IOC)** were converted to SQLite tables. Using Python 3, we were able to easily request, insert, and update relevant information provided by the IOC.

Before using any of the given datasets, we adjusted some of the files for easy processing.

- *summerOly_athletes.csv* - From 1896 to 2024, many **countries changed their official names**. This is reflected within the [Team] column of this CSV. To account for naming discrepancies, the [Name of Country (NOC)] column was utilized. The three-letter NOC code was consistent throughout the sheet, so a dictionary of $\{\text{code} : \text{country}\}$ was used to associate the NOC code with the country name.
- *summerOly_medal_counts.csv* - Due to minor **discrepancies in spacing and naming conventions**, slight changes were made to country names for convenience.

- *summerOly_hosts.csv* - Due to minor **discrepancies in the naming conventions**, slight changes were made to the host names for convenience.
- *summerOly_programs.csv* - For all columns that began with a year, **all empty cells and non-numerical cells were filled with 0**. Those columns otherwise indicated the discontinuation of sports; however, a 0 would likewise indicate that an event is not being offered. **This was more convenient for setting up our SQLite table**.

3 Problem 1: Medal Forecasting

We are tasked with forecasting the total medal counts and gold medal counts for countries in the 2028 Summer Olympics. Given past data since 1896 of country performance in terms of medal counts, broken down into gold, silver, and bronze, as well as athletes in each event, it is reasonable to frame this as a regression problem.

The factors we believe can be used to predict performance are **Host Status**, **Number of Athletes**, and **Past Performance**. Host Status can be directly obtained from the given data. We will also consider if host status has some residual effect on country performance, so if a country is set to host in 4 years or hosted 4 years prior, it is possible this has an effect on performance as that country will put more funds into preparing competitors. Number of Athletes and past performance (in terms of medal counts) can also be directly obtained from the given data. Obviously, there are other factors that can have a major effect such as the amount of money a country puts into training Olympic competitors, but this is data we do not have access to.

If we tried direct regression on the data, we'd run into some interesting issues. By nature, it is impossible for countries to win less than 0 medals at the Games. We are attempting to use regression to predict future performance, and thus forecast the skill of a country's competitors. There is an overwhelming number of zeroes in the dataset, but these zeroes represent countries at very different skill levels. A country could be capable but simply unable to break into 3rd place at any events, or it could be placing last at every event it participates in. This means our data is **censored** with a lower bound of 0, and this is something we must take into account when performing regression.

From our literature review, a Tobit model has been historically used and verified to predict Olympic medal counts. Besides using it to compare with previous models, a Tobit model is also fitting as it assumes the driving mechanism(s) that determine if a country wins medals affects how many medals they win in the same fashion.[2] Therefore, in line with previous research, we will first consider a Tobit model with our predictors.

3.1 Variable Definitions

Table 1: Variable definitions

Variable	Definition
y	Observed Variable
y^*	Latent Variable (on censored data)
$\boldsymbol{X}$	Predictor Data Matrix
$\boldsymbol{\beta}$	Vector of Linear Combination Coefficients for $\boldsymbol{X}$
ϵ	Error term ($\epsilon \sim N(0, \sigma^2)$)
σ	Standard Deviation of Error
σ_n	Random Noise Standard Deviation
σ_u	Unobserved Random Effects Standard Deviation
$L(\beta, \sigma)$	Likelihood Function
$P(A\|B)$	Probability of event A given event B (P is the probability measure)

Φ	Standard Normal Cumulative Density Function (CDF)
ϕ	Standard Normal Probability Density Function (PDF)
$\ell(\beta, \sigma)$	Observed Variable
t_i	T-statistic
$SE(\hat{\beta}_i)$	Standard Error of Fitted Coefficient $\hat{\beta}_i$
μ	Mean for Negative Binomial Distribution
θ	Dispersion Factor for Negative Binomial Distribution
$\boldsymbol{Z}$	Random Effects Design Matrix
$\boldsymbol{b}$	Vector of Linear Combination Coefficients for $\boldsymbol{Z}$
$\hat{y}$	Predicted y value

3.2 General Assumptions

1. **We assume that the provided data is fully accurate to the results of the Olympics each year.**

   Justification: We plan on using regression to predict the 2028 results, and this regression will be usable only if the data we apply it to is accurate.

2. **We assume that country performance can be reasonably estimated using predictive factors associated with Host Status, Number of Athletes, and Past Performance.**

   Justification: We have access to limited data for this problem, and do not have access to (for example) expenditures on Olympic training. Therefore, we should be able to use the data presented to us to predict performance, and these are the main factors that we can extract from the data tables that we think can predict performance.

3. **We assume that the participating countries in the 2028 Summer Olympics are identical to that in the 2024 Olympics.**

   Justification: We do not have data on participating countries in the 2028 Olympics, so we assume all countries that participated in 2024 will also participate (with the exception of Russia, which has had some controversy about participation due to the Ukraine War).

3.3 Tobit Model

The Tobit model accounts for censorship by predicting a latent variable, y^*, instead of directly predicting the observed variable y with the given predictors and previous data. The values y^* and y are related as follows:

$$y = \begin{cases} 0 & \text{if } y^* \leq 0, \\ y^* & \text{if } y^* > 0. \end{cases}$$

The Tobit model attempts to predict the uncensored data itself given only censored data. This fits the framework of Olympic medal counts since, as discussed, zero medal counts can be thought of as censored data. The Tobit model assumes this relationship can be captured from the predictors linearly with the equation

$$y^* = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$$

Where $\boldsymbol{X}$ is a matrix with the predictor data and $\boldsymbol{\beta}$ is a column vector of coefficients (so $\boldsymbol{X}\boldsymbol{\beta}$ is some linear combination of the predictor variables) and ϵ is an error term which is normally distributed with variance σ^2.

This is usually written as $\epsilon \sim N(0, \sigma^2)$. Since our data is limited, we expect that there are discrepancies between what they can predict given the complicated nature of why countries win olympic medals. Generally, ϵ is usually used as a 'noise' term, but in our implementation we extend this to account for unobserved random effects. To do this, we have both σ_n (noise) and σ_u (unobserved effects) that contribute to the variance of ϵ as follows:

$$\sigma^2 = \sigma_n^2 + \sigma_u^2.$$

This ensures that we can capture unobserved effects from factors that we cannot add into linear predictors. From here, our implementation of the Tobit model fits coefficients in $\boldsymbol{\beta}$ based on previous data to be able to predict y^* given predictors $\boldsymbol{X}$.

3.3.1 Maximum Likelihood Estimate (MLE)

Our implementation of the Tobit Model optimizes fit using the **Maximum Likelihood Estimate (MLE)**, which is standard for Tobit models since it can account for both observed and unobserved (censored) parts of the data. Essentially, MLE aims to maximize the likelihood function, which is the probability that y is observed given values of β and σ. Maximizing the likelihood means the fitted coefficients for β and σ yield the largest probability of the actual observed data being observed, which allows us to accurately predict y^* given only predictors $\boldsymbol{X}$.

$$L(\beta, \sigma) = P(y|\beta, \sigma)$$

Across many observations (y_i) this becomes

$$L(\beta, \sigma) = P(y_1, y_2, \dots y_n|\beta, \sigma) = \prod_{i=1}^{n} P(y_i|\beta, \sigma)$$

The likelihood function can then be split up into values where $y_i = 0$ (censored) and $y_i > 0$ (uncensored). The Tobit model generally assumes y^* is normally distributed, which is a restricting assumption but can be altered if necessary. When $y_i = 0$, we want to compute $P(y_i = 0|\beta, \sigma) = P(y^* \leq 0|\beta, \sigma)$, which can be found with the normal **cumulative density function (CDF)**, Φ. When $y_i > 0$, we just want $P(y^* = y_i|\beta, \sigma)$, which can be found with the normal **probability density function (PDF)**, ϕ. Mathematically this becomes

$$\begin{aligned} P(y^* \leq 0|\beta, \sigma) &= \Phi\left(\frac{-X_i\beta}{\sigma}\right) \\ P(y^* = y_i|\beta, \sigma) &= \phi\left(\frac{y_i - X_i\beta}{\sigma}\right) \\ \implies \prod_{i=1}^{n} P(y_i|\beta, \sigma) &= \prod_{y_i=0} \Phi\left(\frac{-X_i\beta}{\sigma}\right) \prod_{y_i>0} \phi\left(\frac{y_i - X_i\beta}{\sigma}\right). \end{aligned}$$

Aiming to maximize, it helps to take the logarithm of the likelihood function. This helps the function in terms of differentiability (to be maximized), and whatever maximum is achieved for the log of the likelihood function will correspond to a maximum of the likelihood function itself.

$$\ell(\beta, \sigma) = \log(L(\beta, \sigma)) = \sum_{y_i>0} \log\left(\phi\left(\frac{y_i - X_i\beta}{\sigma}\right)\right) + \sum_{y_i=0} \log\left(\Phi\left(\frac{0 - X_i\beta}{\sigma}\right)\right)$$

Maximizing this (using various numerical methods) give us fitted parameters β and σ for the Tobit model.

3.3.2 Predictor Analysis and Issues with Tobit

We are given limited data on countries to make this prediction, namely only past performance, number of athletes, and host status. Therefore, it is important that we figure out which are influential/significant in predicting medal count. For each year, we first consider the variables **Host_Status, Num_Athletes, Host_4_Years_Ago, Host_in_4_Years, Past_4_Years_Total_Medals**, and **Past_8_Years_Total_Medals**. The **Host_Status** variables are implemented as binary components. It's reasonable to think that host status can have a delayed or anticipatory effect, as host countries can prepare for hosting by training their athletes and increasing funding to a higher level than usual. We consider past performance as the last two Olympic performances, since these are likely most indicative of future performance rather than performances further in the past.

First, we check for **multicollinearity**. If two predictors are highly correlated, it's not very useful to use them both so we can either combine them or only use one. We check this with a heat map of a correlation matrix between predictors over all country performances. We can see a high correlation between **Past_4_Years_Total_Medals** and **Past_8_Years_Total_Medals** and also **Num_Athletes** in the bottom right 3×3 square. Otherwise, there is not a high correlation with other predictors.

Another thing we must check, especially for the Tobit model, is **homoscedasticity**. The Tobit model relies on the assumption of homoscedasticity, which means that the variance of the residuals is constant, which can be checked in a residual plot as in Figure 1. As we can see, the variance of the residuals is **not** constant, so we have a heteroscedastic error. The reason this is an issue is because after fitting our variance for the error term σ, we are assuming error terms follow a constant variance, however we see here that this is not the case. This is particularly bad for the Tobit model since it can make the probability of censoring (getting 0 medals) very inaccurate.

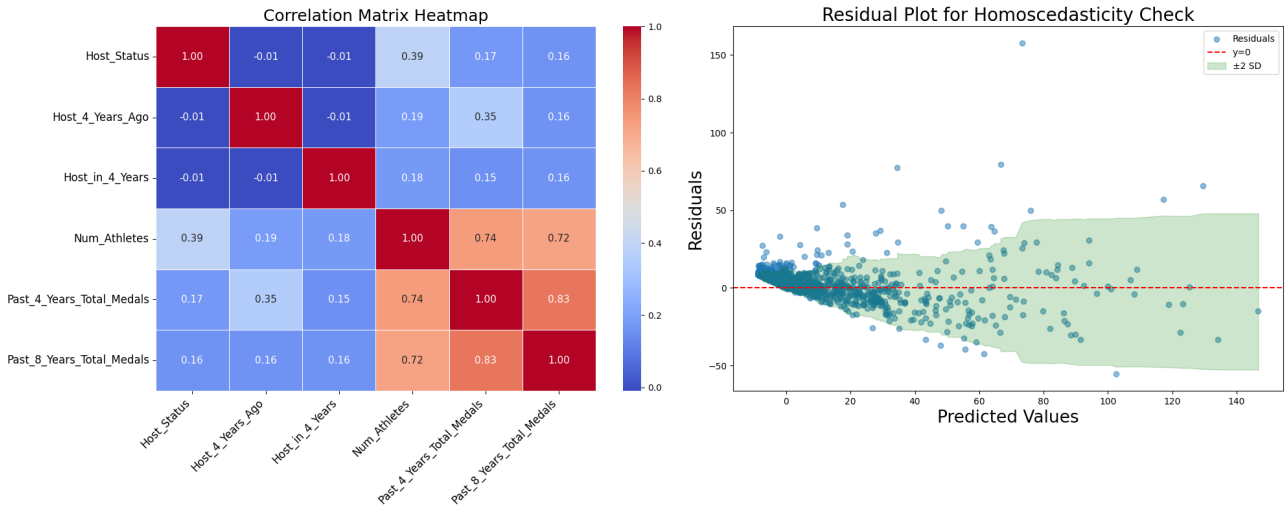


Figure 1: Predictor/Regression Analysis

Further, we perform **regression analysis** to see which predictors are most significant. After fitting, each variable gets a **T-Stat** and **p-value**, defined as

$$t_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Here, β_i is the fitted parameter for predictor i and $SE(\beta_i)$ is the standard error, or an uncertainty metric for β_i (which can be computed in a few ways, we used the covariance matrix).The p-value is calculated using the t-statistic and the degrees of freedom (excess data to predict on), and it is essentially the probability of

that t-statistic being obtained. To determine a statistically significant predictor, it should have a low p-value, generally we want $p \leq .05$. As we can see, the predictors that have this are **Num_Athletes, Host_in_4_Years**, and **Past_4_Years_Total_Medals**. From this, we can reasonably eliminate **Past_8_Years_Total_Medals** due to multicollinearity and low statistical significance and **Host_4_Years_Ago** for low statistical significance. Though **Host_Status** isn't statistically significant in this analysis, we may still consider it in the model improvement to see if anything has changed.

Variable	Coefficient	T-Stat	P-Value
Host_Status	-4.370	-0.642	0.521
Num_Athletes	18.356	17.246	0.00
Host_4_Years_Ago	0.092	0.000	0.99
Host_in_4_Years	-13.225	-12.608	0.00
Past_4_Years_Total_Medals	3.733	3.683	0.00023
Past_8_Years_Total_Medals	0.453	0.003	.997

Table 2: Regression Analysis Results

3.4 Mixed Linear Models (MLMs) in a Hierarchical Framework

We intend to refine the Tobit model by attempting to account for differing levels of variance. The purpose of the random variable is to account for unobserved effects, and there's no reason to assume this is the same across all countries (in fact, that is quite a bad assumption). We decided to fix this by splitting countries into **clusters** based on general performance, including separate random variables for each cluster as well as each individual country. This is done in attempt to capture important effects on a country-level and cluster-level scale. These underlying effects could possibly be socioeconomic or geographic, for example low performance can be due to countries that do not have the money to put towards Olympic training.

3.4.1 Clusters with DBSCAN

First, to form clusters, we used a **Density-Based Spatial Clustering of Applications with Noise (DBSCAN)** model. DBSCAN is a useful method for us since it specifically designates outliers (noise) as high performing countries are in this data. To get a general metric for country performance, we define 'Points' as a weighted sum of the number of Gold, Silver, and Bronze medals, with weights 6, 3, and 2 respectively. All weightings are arbitrary—different reputable sources use different weighting systems—but we decided to use this weighting because, to some extent, it has a natural justification, as it follows Zipf's law. That is,

$$Points = 6 \cdot Gold + 3 \cdot Silver + 2 \cdot Bronze.$$

Then, countries are ranked by their average number of points over the years they competed. The DBSCAN algorithm takes two inputs: *eps* and *min_points*. *min_points* is the minimum number of points that can be in a cluster, and *eps* is a distance threshold to define a neighborhood around a 'core points'. A core point has at least *min_points* within distance *eps* of it, and boundary points to each cluster are points within *eps* of a core point. DBSCAN algorithmically goes through all points and forms clusters.

We wanted to split countries into clusters large enough to perform regression on but also have enough clusters to model some inter-cluster unobserved effect difference. In the end, we decided on 3 clusters (including outliers). Our code took $min\_points = 8$ and with a target number of (non-outlier) clusters of 2, we found a fitting *eps* value through iteration shown in Figure 2.

After finding the clusters, we applied a Mixed Effects Linear Model to each cluster, taking special care of Cluster 1 (which included all of the countries that get close to 0 medals).

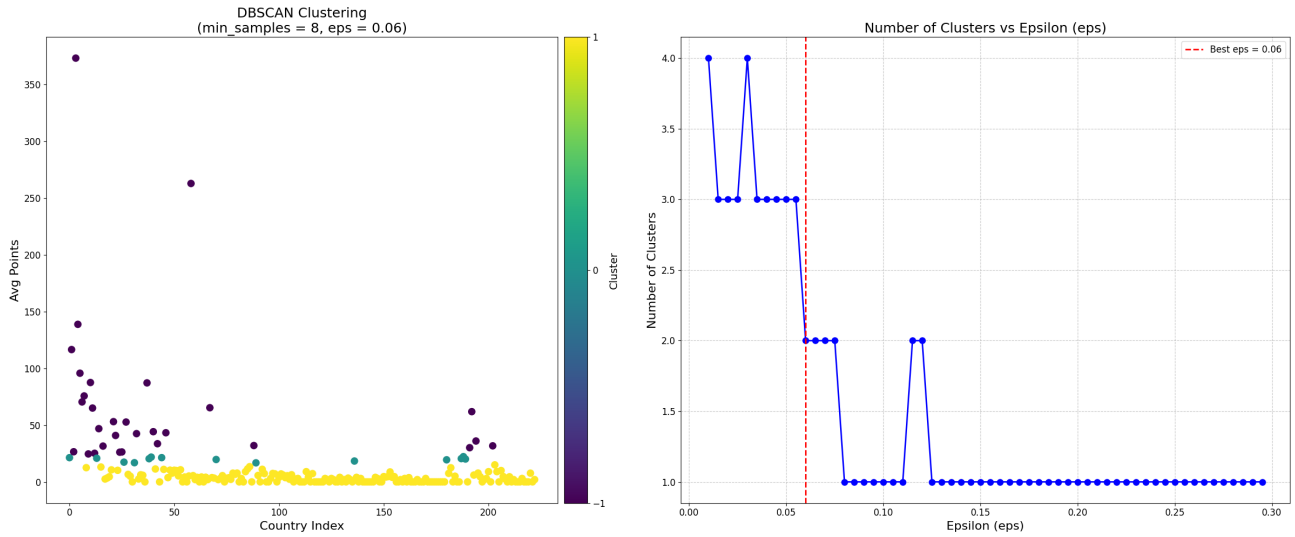


Figure 2: DBSCAN Cluster Results

3.4.2 Mixed Linear Models and Hurdle Model

We will first describe the model for clusters 0 and -1, then for cluster 1. The Tobit model was a continuous regression model, however since medal counts are discrete we chose to use a Negative Binomial Regression. A Negative Binomial distribution is chosen over a Poisson since the Negative Binomial can account for cases where the variance is greater than the meain (Poisson is too restrictive). For countries in these clusters, we assumed there was close to negligible chance of them getting 0 medals. The Mixed Effects Model takes the following form:

$$y_{ij} \sim \text{NBinom}(\mu_{ij}, \theta)$$
$$\mu_{ij} = \exp(\boldsymbol{X}_{ij}\boldsymbol{\beta}_j + \boldsymbol{Z}_{ij}\boldsymbol{b}_j)$$

Here i is the index for individual country and j is the index for cluster. θ is a dispersion factor for the negative binomial distribution that is also fitted. Now, $\boldsymbol{Z}_{ij}\boldsymbol{b}_j$ is implementing group-wise random effects for each parameter in attempt to reduce heteroscedasticity (though this model suffers less from it since we are using Negative Binomial). $\boldsymbol{b}_j$ is a vector of random coefficients to apply to each cluster, and these random terms are normally distributed. The matrix $\boldsymbol{Z}_{ij}$ is the random-effects design matrix, with predictor inputs that get multiplied with the random effects variables in $\boldsymbol{b}_j$ (so $\boldsymbol{Z}_{ij}$ is similar to $\boldsymbol{X}_{ij}$ in this manner). This is a 'mixed' effects model as $\boldsymbol{X}_{ij}\boldsymbol{\beta}_j$ is the fixed part and $\boldsymbol{Z}_{ij}\boldsymbol{b}_j$ is the random part of the model. Cluster 0 and -1 were fitted this way, with a slightly different process for Cluster 1.

The purpose of choosing the Tobit model was to account for censorship, but another way to deal with this is a Hurdle model. It is very similar, first using binary regression (Probit) to determine the probability a country gets 0 medals, and then a Negative Binomial model as previously described for the positive count model. The random effects modeling is identical for the Probit regression, but instead a different link function is used. This is implemented mathematically as follows:

$$P(Y_{ij} = 1 | \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}) = \Phi(\boldsymbol{X}_{ij}\boldsymbol{\beta}_j + \boldsymbol{Z}_{ij}\boldsymbol{b}_j)$$

The value $P(Y_{ij} = 1)$ is what we are trying to predict in the binary regression (the probability a country gets any medals at all), and we take the cumulative normal distribution of the linear predictor for the probit model. In these mixed linear effects models, it's useful to think of a link function that relates the linear predictor to the

desired value. In the probit model, the link is the inverse CDF, and in the negative binomial, it is the logarithm. That is, we have

$$\Phi^{-1}(P(Y_{ij} = 1)) = \boldsymbol{X}_{ij}\boldsymbol{\beta}_j + \boldsymbol{Z}_{ij}\boldsymbol{b}_j$$
$$\log(\mu_{ij}) = \boldsymbol{X}_{ij}\boldsymbol{\beta}_j + \boldsymbol{Z}_{ij}\boldsymbol{b}_j$$

Now the right side is the same, as the base for these models is a linear predictor, and changing the link function allows us to model different scenarios. For Cluster -1, we combine the probit binary regression and the negative binomial as follows:

$$\text{Predicted Number of Medals} = (1 - P(Y_{ij} = 0)) \cdot \hat{y}_{ij}$$

Where for clusters 0 and -1 the predicted number of medals is just $\hat{y}_{ij}$. The optimization method here was also MLE, though a different one could have led to a better fitting model.

3.4.3 Results, Model Fit, and Uncertainty

We decided to use **Akaike Information Criterion (AIC)** and **Bayesian Information Criterion (BIC)** as well as the log likelihood value to determine the quality of our model's fit to the data. Both AIC and BIC are based on the log likelihood as well as number of parameters and sample size. Generally lower AIC/BIC and higher log likelihood is desired for a well fitted model. In fitting the mixed effects models, we experimented with different methods to determine one that would optimize these metrics of fit. The one that worked the best was Powell's method, which is just one of many optimization methods but happened to fit best. However, it may be due to the nature of the data that MLE never quite converged and our fit metrics ended up subpar. Checking again for heteroscedasticity showed us that this model did not improve it as much either, however this model is less susceptible to the inaccuracies Tobit would face from heteroscedasticity.

For each of the cluster models, here are our final results (for total medals): Originally, even with the best

Metric	**Cluster 1**	**Cluster 0**	**Cluster -1**
Zero Model (Cluster 1 only)			
Log-likelihood	-746.77	-	-
AIC	1523.55	-	-
BIC	1609.37	-	-
Positive Model / Full Model			
Log-likelihood	-1018.16	-590.75	-1083.85
AIC	2058.32	1211.51	2195.71
BIC	2105.87	1262.55	2259.12

Table 3: Model Fit Statistics by Cluster

method (Powell's), cluster -1 had exceptionally bad metrics (likely since it is a group of outliers). To fix this, we weighted all of the predictive factors by the variance in an attempt to reduce heteroscedasticity and fit better when optimizing with MLE. This worked, cutting the AIC/BIC factors in half and increasing the log-likelihood level to one similar to the other models. Generally, our model accounts for random effects on multiple scales and attempts to combat heteroscedasticity, and we have chosen the best methods to fit the data (despite convergence issues). To predict the number of athletes for each country in 2028, we used a simple linear regression. Finally, we applied this model to Total Medal count, Gold, Silver, and Bronze. We will provide the top 7 countries ordered by Gold medals, but our code predicts full Olympic results.

Likely in line with the fit metrics not being the best, there is a large confidence interval margin. However, these results do have some reasonable conclusions.

Rank	Country	Gold	Silver	Bronze	Total
1	United States	48 [33, 63]	43 [32, 54]	45 [13, 77]	136 [108, 164]
2	Germany	28 [13, 43]	23 [12, 34]	22 [0, 44]	73 [45, 101]
3	China	25 [10, 40]	18 [7, 29]	27 [0, 54]	70 [42, 98]
4	Great Britain	17 [2, 32]	19 [8, 30]	17 [0, 38]	53 [25, 81]
5	Japan	16 [1, 31]	9 [0, 20]	13 [0, 35]	38 [10, 38]
6	Australia	13 [0, 28]	17 [6, 28]	19 [0, 41]	49 [21, 77]
7	Italy	11 [0, 26]	10 [0, 21]	20 [0, 43]	41 [13, 69]

Table 4: Predicted Medal Counts for Top 7 Gold Medal Countries in 2028 Olympics (with 95% Confidence Intervals)

We are also tasked with looking for countries that have improved or gotten worse since the 2024 Olympics. We did this by looking at rank change, setting a minimum medal change threshold of 3 so that it wasn't just countries that changed a lot even though they all got 0 or low counts.

Country	Rank 2024	Gold 2024	Rank 2028	Gold 2028	Rank Change	Medal Change
Most Improved						
Jamaica	43	1	16	7	+27	+6
Germany	10	12	2	28	+8	+16
Ukraine	23	3	15	7	+8	+4
Spain	15	5	8	11	+7	+6
Most Declined						
Canada	12	9	111	0	-99	-9
Netherlands	6	15	37	2	-31	-13
South Korea	8	13	18	6	-10	-7
France	5	16	11	10	-6	-6

Table 5: Countries with Most Significant Changes in Olympic Performance (2024 vs 2028)

3.5 Projections for Countries Yet to Earn Medals

Our model accounted for countries that generally don't medal through clustering and a specific binary regression model to determine if those countries do medal in a certain year or not. Our model predicted these values as stated previously in the paper as

$$\text{Predicted Number of Medals} = (1 - P(Y_{ij} = 0)) \cdot \hat{y}_{ij}$$

Here $\hat{y}_{ij}$ is the number of medals they would earn if they medalled at all, and it is multiplied by the probability of earning a medal. We looked at the list of countries that have never medalled in Olympic history and compared it to our predictions for 2028. We assigned odds to these estimates as the probability these countries medalled at all from our regression (the odds are $P(Y_{ij} = 1)$). We found that two countries, **Angola** and **Saar**, are predicted to win 1 medal in the 2028 Olympics and have never won a medal previously, with odds 32.54% and 40.08% respectively.

3.6 Analyzing Impacts of Events and Sports

Through the information in the data sets, **two deciding factors** in the performance of a country at the Games were the **amount of medals** they obtained and the **distribution of medals** among Gold, Silver, and Bronze.

However, before analyzing the impacts of events and sports on the performance of a country at the Games, **a lot of processing** wasa done to create viable connections between the datasets.

3.6.1 Data Processing (Ceiling-Proportion Approximation)

1. The *medal_counts.csv* file provided the total amount of medals obtained by each country (that wins at least 1 medal) and the distribution of those medals during all the Games. However, **there is no clear connection to the distribution of those medals to sports and events**.
2. By using *athletes.csv* to process specific athletes, this CSV file could be used to total the amount of medals obtained in a specific sport or event (with other parameters like country, year, etc). This was beneficial as it created a distribution of medals to events and sports. However, this created another issue. **There was an over counting of medals due to overlapping events**. Due to the existence of team events, even if many athletes obtain a medal, it may only be 1 medal for the country. To account for this, *programs.csv* provided helpful data.
3. *programs.csv* contains information regarding the number of sports per discipline. This meant that there's **a known ceiling to the amount of medals that could be obtained per sport**.

Since there is a known ceiling, the next step was **creating a proportion that could multiply with the ceiling to approximate the medal distribution in sports**. Going back to *athletes.csv*, we were able to find the amount of athletes from a country that obtained a medal in a specific sport. By removing one of the parameters, we could also find the amount of athletes that obtained a medal in a specific sport. These two different values (different in country specification), allowed us to create **a proportion between a country's total athlete medal count in a sport to the total athlete medal count in a sport**.

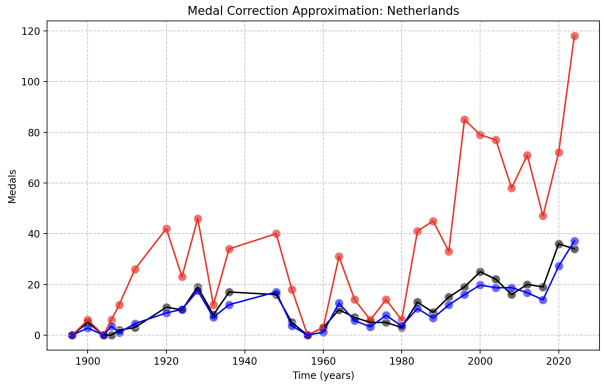
$$\text{Ceiling-Proportion Approximation} = \frac{\text{Country's total athlete medal count in a sport}}{\text{Total athlete medal count in a sport}} \cdot 3 \cdot \text{Total Events per Sport}$$

This **Ceiling-Proportion Approximation** would yield an estimate of the amount of medals in a given sport (note that we multiply by 3 as there are typically 3 medals to be obtained per event). This gave us two different approximations for medal distribution in sports: the distribution solely from the amount of athletes from a country that obtained a medal in a specific sport and the distribution based on the ceiling multiped by a related proportion. To **test the accuracy** of the two different approximations, we isolated the sports in each year, then summed the applied approximation of medal counts in every sport offered to get a total medal count. Then, we **analyzed the accuracy of the total medal count approximations over time** on the United States and the Netherlands. This can be seen in Figures 3a and 3b.

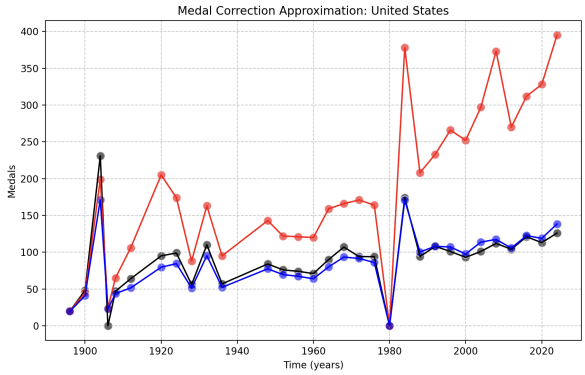
The figures demonstrated that the approximation of total medal counts using the ceiling-proportion approximation was **much more accurate** than just using the total medals obtained by all of a country's athletes in a sport. We verified this by taking the **average percent error** of both approximations and comparing them. In the Netherlands, the **ceiling-proportion approximation** had a percent error of **20%** while the **athlete total approximation** returned **198%**. Similarly for the United States, our program returned **10% and 107% error respectively for the ceiling-proportion and athlete-total approximations**.

By using the ceiling-proportion approximations for total medal count that were built on approximations for medal distribution in sports, we could get **reasonable estimates for the medal counts in various sports and disciplines**. However, there was one more issue before we could start analyzing the distribution of medals in sports. Not all medals are the same. This approximation gave us estimates for the total amount of medals, but not for the distribution among Gold, Silver, and Bronze. Luckily, we were able to apply a similar approximation as before:

$$\text{Ceiling-Proportion Approximation} = \frac{\text{Country's total athlete } \textbf{specific} \text{ medal count in a sport}}{\text{Total athlete } \textbf{specific} \text{ medal count in a sport}} \cdot \text{Total Events per Sport}$$



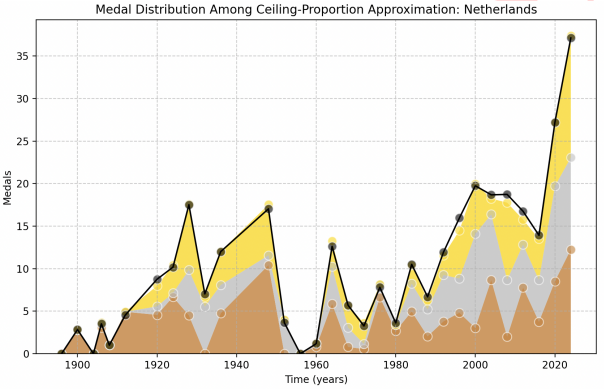
(a) Total medal approximations in Netherlands compared to actual medal counts



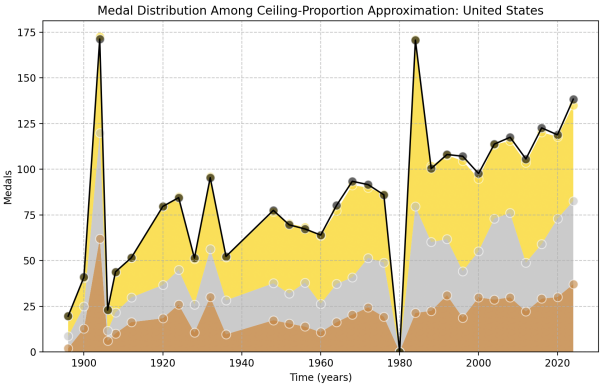
(b) Total medal approximations in United States compared to actual medal counts

Figure 3: Black Marks: The **actual** total medal counts, Blue Marks: **ceiling-proportion** approximation, Red Marks: **athlete total** approximation .

Notice that in this Ceiling-Proportion Approximation, we removed the scalar of 3. This scalar acted to account for all medal types earlier, but now that we are giving a **specific medal**, we no longer need to multiply by a factor of 3. Instead, we will apply this Ceiling-Proportion Approximation to each medal type. With three approximations for Gold, Silver, and Bronze, we can build a new approximation for the total medal count. This time giving us the distribution of medal types. We can see this in Figures 4a and 4b.



(a) Total medal approximations in Netherlands compared to actual medal counts



(b) Total medal approximations in United States compared to actual medal counts

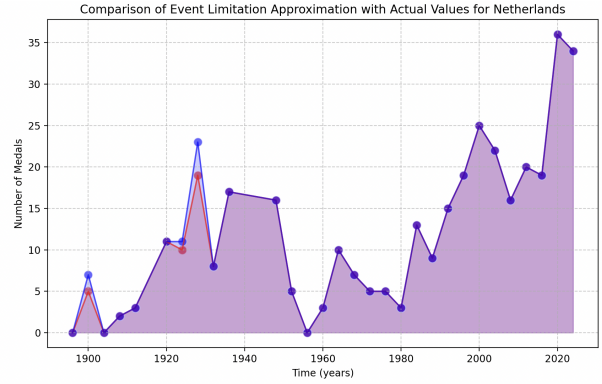
Figure 4: Black line: **Ceiling-Proportion Approximation from Figure 3**, Brown Region: The **Bronze** medal count estimate, Grey Region: **Silver** medal count estimate, Yellow Region: **Gold** medal count estimate

These approximations for total medal counts (with respect to our initial Ceiling-Proportion Approximations from Figure 1) were promising as there was **only 2.8% and 1.1% average percent error** respectively for the Netherlands and United States. However, it's important to keep in mind that this approximation is still problematic as it **creates the assumption that if the medal distributions sum to a reasonably accurate total medal count, then the medal distribution will also be accurate**.

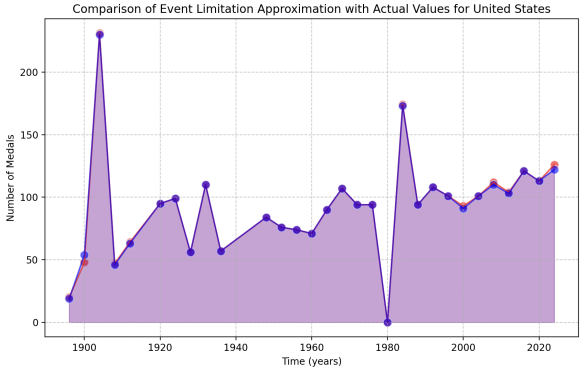
3.6.2 Data Processing (Event Limitation Approximation)

While the Ceiling-Proportion Approximation provided a fairly reasonable estimation of the medal distribution among sports/disciplines, **we lacked a method of associating sports/disciplines to events**. This inspired our Event Limitation Approximation approach. This approximation works as **in almost all cases, an event only has 1 medalist in Gold, Silver, and Bronze**. By creating a Python dictionary to store all isolated unique events that a country obtained a medal in, the length of that dictionary would tell us how many Gold, Silver, or Bronze medals they received and its distribution among sports and events.

This method parses through all of the data within *athletes.csv* and collects every single entry that has a specified year. Then, we **reduced the data** to only give us the country code, sport, event, and medal type. A **dictionary proved useful** here to **distinguish unique events**. We required three separate dictionaries for each medal type as it was possible for a key of (country code, sport, event) to be associated to Bronze, Silver, or Gold (which is not allowed by Python dictionaries). By searching each dictionary for a unique country code (ie. USA for United States), we could create subsets per country that allow us to know how many Gold, Silver, and Bronze medals they got as well as in what events. This **satisfied our need for associating medals to events**. To test the accuracy of this approximation, we ran our code for the United States and Netherlands again. The results are shown in Figures 5a and 5b.



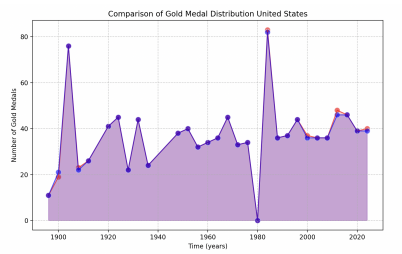
(a) Total medal approximations in Netherlands compared to actual medal counts



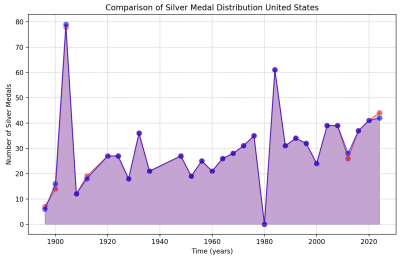
(b) Total medal approximations in United States compared to actual medal counts

Figure 5: Red: Actual values given by IOC, Blue: Our Event Limitation Approximation

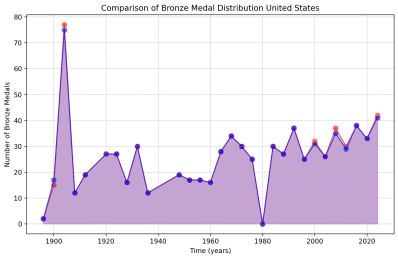
This was **very accurate** with **average percent errors of 2.4% and 1% for the Netherlands and United States** respectively. Now, we want to compare our distribution of medal types with the IOC's distribution. We can see this below in Figures 6a, 6b, and 6c.



(a) Gold Medal Distribution Comparison in United States



(b) Silver Medal Distribution Comparison in United States



(c) Bronze Medal Distribution Comparison in United States

Figure 6: Red: Actual values given by IOC, Blue: Our Event Limitation Approximation

This was **very promising** as we only found **0.85%, 1.6%, and 1% error** for gold distribution, silver distribution, and bronze distribution respectively.

3.6.3 Selecting a Processing Method

When **choosing** between our Ceiling-Proportion Approximation and Event Limitation Approximation, **we considered three factors**:

1. **Accuracy of Total Amount of Medals**: Comparing the percent errors from Figure 1 and Figure 3, the Ceiling-Proportion Approximation found **percent errors of 20% and 10%** while our Event Limitation Approximation found **percent errors of 2.4% and 1%** for the Netherlands and United States respectively.

2. **Accuracy of Medal Distribution into Gold, Silver, and Bronze**: As described earlier, there is a **huge flaw in our Ceiling-Proportion Approximation** for medal distribution as **its an approximation of an approximation**. This assumes the behavior of various factors like team events that we're simply unable to account for. Alternatively, our **Event Limitation Approximation found percent errors lower and on-par** with the percent errors from our Ceiling-Proportion Approximation. However, the Event Limitation Approximation found these percent errors with respect to the IOC data. **This removed all assumptions made when comparing medal distribution while maintaining a low percent error.**

3. **Computation Time**: The Ceiling-Proportion Approximation takes approximately **3-5 minutes per country** while our Event Limitation Approximation takes approximately **3-5 seconds per country**. This was another clear indicator to choose the Event Limitation Approximation as it **allowed us to process data for large groups of countries**.

Our **Event Limitation Approximation out-performed our Ceiling-Proportion Approximation** in all three factors that we considered for choosing a processing method. For those reasons, we decided to use the Event Limitation Approximation.

3.6.4 Event Impact: Singular Value Decomposition (SVD)

We will look at the data for all events competed in 2024, and create a matrix with medal counts for each country by event. Then, we perform **Singular Value Decomposition (SVD)** on this matrix A, represented as

$$A = U\Sigma V^T$$

Where Σ is a diagonal matrix with the singular values (square root of the eigenvalues of A). The magnitude of these eigenvalues indicates the strength of the scaling in the direction of their associated column vectors. The singular values in the diagonal matrix Σ determine how much each corresponding singular vector contributes to the overall transformation of the matrix, with larger singular values indicating stronger influence in the corresponding direction.

Based on the data from the United States in 2024, we were able to isolate the Top 10 influential events in medal counts. And by applying the SVD to further countries, we could understand how each event plays a role in the medal counts of all countries competing in the Games.

3.6.5 Analyzing Impact of Sports

It is important to note that the naming systems in the files are inconsistent: we analyzed performances in categories classified as "sports" in *athletes.csv*, even though this category includes both sports and disciplines as designated in *programs.csv*.

Rank	Singular Value	Event
1	3838.9080	Men's Greco-Roman 97kg
2	2664.5821	Men's Greco-Roman 77kg
3	2397.0722	Men's Greco-Roman 67kg
4	2042.4828	Men's Freestyle 57kg
5	1903.1633	Men's Freestyle 65kg
6	1819.0644	Women's 20km Race Walk
7	1666.6152	Marathon Race Walk Relay Mixed
8	1533.8462	Women's 100m
9	1468.2113	Women's 200m
10	1296.5435	Men's 100m

Table 6: Ranked Singular Values and Associated Events

Here, we are tasked with understanding what sports are most important to various countries. We deviate from 4.6.4 as we will now consider the previous ranking point system instead of total medals. Each gold medal is worth 6 points, each silver medal is worth 3 points, and each bronze medal is worth 2 points. By totaling up the points, we can weigh the effect of the medal distribution on the total medal count. To quantify the impact of sports, we will consider the following three factors:

1. Sport-Point Efficiency: A proportion between the points a country obtains in one sport and how many athletes from that same country competed in that sport.
2. Sport-Point Share: A proportion between how many points a country obtains in one sport and how many points a country obtains at one Olympic Game.
3. Revealed Comparative Advantage (RCA) Index: While the RCA Index is typically applied to reveal a country's advantage in producing goods and services, we can extend this to the Games. The RCA Index here will tell us if a country has an advantage in a sport when compared with other countries' data given by the IOC.

Variable	Definition	Unit
P_{cs}	Number of points won by country c in sport s.	Count
A_{cs}	Number of athletes by country c in sport s.	Count
$\sum_s P_{cs}$	Total points won in the Games by country c.	Count
$\sum_c P_{cs}$	Total points won in the Games in sport s.	Count
$\sum_{c,s} P_{cs}$	Total points won in the Games across all countries and all sports.	Count
$RCAI_{cs}$	Revealed Comparative Advantage Index by country c in sport s.	Unitless (Index)
SPE	Sport-Point Efficiency.	Unitless
SPS_c	Sport-Point Share for country c.	Unitless

Table 7: Variable Definitions

We will apply each of these indicators of a sport's importance to the United States and Netherlands in the most recent Olympic game: Paris Olympics, 2024. Additionally, since there were too many unique "sports", we will only consider the top 10 per country.

1. Sport-Points Efficiency

$$SPE = \frac{P_{cs}}{A_{cs}}$$

2. Sport-Points Share

$$SPS_c = \frac{P_{cs}}{\sum_c P_{cs}}$$

3. Revealed Comparative Advantage (RCA)

$$RCAI_{cs} = \frac{\frac{P_{cs}}{\sum_s P_{cs}}}{\frac{\sum_c P_{cs}}{\sum_{c,s} P_{cs}}} = SPS_c \frac{\sum_{c,s} P_{cs}}{\sum_s P_{cs}}$$

Sport	P_{cs}	A_{cs}	SPE
Weightlifting	8.00	5	1.600
Wrestling	21.00	16	1.313
Athletics	164.34	126	1.304
Surfing	6.00	5	1.200
Basketball	25.03	24	1.043
Shooting	16.67	17	0.980
Archery	3.50	4	0.875
Golf	6.00	7	0.857
Fencing	14.00	20	0.700
Skateboarding	8.00	12	0.667

Table 8: Top 10 SPE Values for Sports in the United States

Sport	P_{cs}	A_{cs}	SPE
Sailing	15.71	11	1.429
Rowing	32.96	33	0.999
Athletics	43.3	46	0.941
Equestrian	0.8	12	0.067
Volleyball	0.0	13	0.000
Triathlon	0.0	4	0.000
Tennis	0.0	6	0
Handball	0	10	0
Judo	0	10	0
Archery	0	4	0

Table 9: Top 10 SPE Values for Sports in the Netherlands

Sport	P_{cs}	$\sum_c P_{cs}$	SPS_c
Basketball	25	44	0.57
Athletics	164	528	0.31
Surfing	6	22	0.27
Golf	6	22	0.27
Football	6	22	0.26
Aquatics	137	539	0.25
Volleyball	10	44	22.7
Gymnastics	38	198	0.19
Skateboarding	8	44	0.18
Tennis	6.25	55	0.11

Table 10: Top 10 SPS_c Values for Sports in the United States

Sport	P_{cs}	$\sum_c P_{cs}$	SPS_c
Rowing	32.96	154	0.214
Sailing	15.71	110	0.143
Cycling	30.64	242	0.127
Athletics	43.30	528	0.082
Basketball	3.00	44	0.068
Aquatics	16.51	539	0.031
Equestrian	0.80	66	0.012
————	—	—	0
————	—	—	0
————	—	—	0

Table 11: Top 10 SPS_c Values for Sports in the Netherlands

Sport	SPS_c	$\sum_{c,s} P_{cs}$	$\sum_s P_{cs}$	$RCAI_{cs}$
Basketball	0.57	3619	525	3.92
Athletics	0.31	3619	525	2.15
Surfing	0.27	3619	525	1.9
Golf	0.27	3619	525	1.9
Football	0.26	3619	525	1.79
Aquatics	0.25	3619	525	1.75
Volleyball	22.7	3619	525	1.57
Gymnastics	0.19	3619	525	1.30
Skateboarding	0.18	3619	525	1.25
Tennis	0.11	3619	525	0.78

Table 12: Top 10 $RCAI_{cs}$ Values for Sports in the United States

Sport	SPS_c	$\sum_{c,s} P_{cs}$	$\sum_s P_{cs}$	$RCAI_{cs}$
Rowing	0.214	3619	143	5.42
Sailing	0.143	3619	143	3.62
Cycling	0.127	3619	143	3.21
Athletics	0.082	3619	143	2.08
Basketball	0.068	3619	143	1.73
Aquatics	0.031	3619	143	0.78
Equestrian	0.012	3619	143	0.31
————	—	—	—	0
————	—	—	—	0
————	—	—	—	0

Table 13: Top 10 $RCAI_{cs}$ Values for Sports in the Netherlands

Since SPS_c is a building block of $RCAI_{cs}$, we chose to only consider the data from the top 10 SPE and top 10 $RCAI_{cs}$.

SPE by definition tells us the return on investment. If we expect each athlete to contribute at least 1 ranking

point, then $SPE > 1$, tells us that there were more points won in a sport than the amount of athletes. This means that for a lower use of resources on athlete expenditures and training, countries get a higher return of points. This may suggest to countries the importance of sports that have a high return of ranking points. So, for the United States, they may consider investing more in their Top 5 in Table 8. And for the Netherlands, they may consider investing more in Sailing.

If $RCAI_{cs} > 1$, then that sport is considered to have a competitive advantage against the other countries competing in the same sports. This value is integral for determining the importance of sports as it allows a country to know what sports they are outperforming others in. For the United States, this means almost all of the sports in Table 12. While for the Netherlands, this would be their Top 5 within Table 13.

3.6.6 Home Country Effects

It's a well-known fact that the host country often has a competitive advantage whether it's due to psychological factors or a country's ability to select sports (like in the Games). We can see preliminary evidence of a home country advantage just by looking at the increase in medal counts for all host countries.

Table 14: Δ Total Medal Count between host year and previous competition year for last 6 Host Countries

Host	Year	Δ Total Medal Count
France	2024	31
Japan	2020	17
Brazil	2016	2
Great Britain	2012	14
China	2008	37
Greece	2004	3

Now that we know this hosting advantage is present at the Games, we want to understand how the events chosen affect this advantage. According to the IOC, host countries have only been allowed to propose (influence the choosing) of events since the 2020 Tokyo Olympics. This means that to understand how the events chosen impact home country results, we are limited to data from the 2020 Tokyo Olympics and 2024 Paris Olympics.

By applying the RCA index from 3.6.5 on each unique event offered to the host country each year, we could understand the influence of events that the host country selected. Again, an $RCAI_{cs} > 1$ would indicate that the event affects results more than they would for a majority of the other competing countries. However, we could not apply this in the given time.

4 Problem 2: Detecting and Utilizing the Great Coach Effect

4.1 Detecting the Great Coach Effect

Our primary consideration when developing a model to detect great coach effects was the fact that we had very limited data: the only accessible data that was relevant to our model were the point distributions by nation and by sport, which we found by using the event limit approximation method from 4.6.2.

Many other factors—such as outstanding individual athletes or increased government investment in athletic programs—could affect performances, and because of our limited data, it is impossible to show that any changes in the point distributions for a given country and sport are directly due to the influence of a great coach, as they could be (and likely are) results of a combination of factors. In other words, the most we could accomplish with the provided data was finding possible instances of the great coach effect. However, we can examine the results of this model in the context of the history of great coaches to see how accurate it is.

4.1.1 Time Series Analysis Setup

The general strategy behind our model to detect potential great coach effects is identifying a given country's performances in a given sport that are anomalously successful in the context of that country's historical performances in the same sport. To do this, we applied time series analysis to the data.

A time series is a series of data that are equally spaced apart. Aside from a few notable exceptions—the 1916, 1940, and 1944 Games were canceled because of the World Wars—the regular four-year timing of the Games means that performances at the Games are suitable for time series analysis.

Many countries have been competing in the same sport for well over a century. Over this time, the influence of individual great coaches is greatly outweighed by a multitude of other factors. So, if we were to calculate the z-scores of each performance in the context of the country's entire historical performance, the z-scores would be less meaningful. To more accurately interpret variations, we needed to calculate the z-scores of each performance in only the context of the Games immediately preceding and succeeding it. For this, we used a rolling window of four Games at a time (more detail provided in the outline).

4.1.2 Variable Definitions

Table 14: Variable Definitions

Variable	Definition	Formula
N	The total number of Games a given country has competed in for a given sport	
L	Number of Games included in the segment for a given country and sport	
P_s	Number of points earned in the segment for a given country and sport	
μ	Population mean points for a given country and sport over all Games	$\frac{1}{N}\sum_{i=1}^{N} x_i$
σ	Population standard deviation of points for a given country and sport over all Games	$\sqrt{\frac{1}{N}\sum_{i=k}^{N}(x_i-\mu)^2}$
$\overline{x}$	Sample mean points for a given country and sport over a rolling set of four Games	$\frac{1}{4}\sum_{i=k}^{k+3} x_i$, where $k \in \{1, 2, ..., N-3\}$
s	Sample standard deviation of points for a given country and sport over a rolling set of four Games	$\sqrt{\frac{1}{4}\sum_{i=k}^{k+3}(x_i-\overline{x})^2}$, where $k \in \{1, 2, ..., N-3\}$
C	Relative contribution of great coaches for a given country and sport	$(P_s - \mu \cdot L) \cdot 100\%$
ΔP_{norm}	Normalized relative contribution of great coaches for a given country and sport	$\frac{C}{L}$

For a given country and sport, this was the outline of the model:

1. Calculate $\overline{x}$ and s in a rolling window of four Games at a time over the history of the country's participation in that sport.
    (a) E.g. for a country that has participated in a sport since 1896, the windows for that country-sport pair would start from [1896, 1900, 1904, 1908], then [1900, 1904, 1908, 1912], and so on. The means and sample standard deviations of the country's points in each of these sets of years are calculated.

2. When the difference in points between performances at consecutive games is at least $2s$, call these positive change points.
3. When the difference in points between consecutive games is less than or equal to $-3s$, call these negative change points.
4. Let segments be the years between positive and negative change points.
5. If a positive change point is followed by another positive change point before a negative change point appears, combine the segments.
6. For each segment, calculate the z-score of $\overline{x}$ compared to the population distribution of points: $\frac{\overline{x}-\mu}{\sigma}$.
7. Choose segments with $\overline{x}$ that have z-scores greater than or equal to 1.5—these indicate the great coach effect.
8. For each of these chosen segments, calculate the coefficient of variation: $\frac{s}{\overline{x}}$.

4.1.3 Results and Notable Findings

It was impractical to apply this analysis to all pairs of countries and sports, so we chose the 20 countries with the highest points over the history of the Games. Though many great coaches have coached teams outside of these top 20 countries, it is reasonable to assume that many of these top-performing countries have achieved such success at least partly due to the great coach effect, making it more likely that our model detects instances of the effect for these countries' performances.

The table below shows the top 10 segments, ranked by order of decreasing z-score.

Table 16: Analysis of Significant Performance Changes in Olympic Sports

Year		Z-Score	Coefficient	Country (Sport)
Start	End		of Variation	
1972	1976	2.497	0.222	Romania (Handball)
2004	2008	2.284	0.286	Canada (Trampolining)
1996	2000	2.264	0.353	Russia (Diving)
2004	2008	2.207	0.622	China (Trampolining)
2000	2004	2.167	0.667	Russia (Basketball)
2000	2004	2.146	0.374	Russia (Weightlifting)
1984	2004	2.143	0.547	South Korea (Handball)
1996	2016	2.100	0.286	Russia (Modern Pentathlon)
1908	1912	2.076	0.766	Netherlands (Fencing)
1996	2000	2.070	0.182	Australia (Basketball)

Here, the start year is the year before the potential great coach effect, and the end year is the last year of the potential great coach effect. Matching with known information about coaches, many of these results line up with the leadership of coaches who may be considered "great": the 1976 Romanian handball team was coached by Radu Voina, who won the world championship and three Olympic medals himself;[3] the 2008 Canadian trampolining team was coached by David Ross, whose club athletes have won seven Olympic medals;[4] and so on. Many of these coaches had very successful coaching careers and may be considered particularly great coaches, even among the exceedingly high standards that Olympic coaches are held to.

Most of these great coach effects last only for one Games, but this makes sense when considering the fact that it is common for coaches to switch teams: in a particularly notable case, Lang Ping, the "great" volleyball

coach, coached the Chinese team to silver in 1996, the American team to silver in 2008, and the Chinese team to gold in 2016.

This is why we did not use the coefficient of variation as a variable in determining whether a segment of successful performances was likely to be due to the great coach effect: it would disqualify some segments with high coefficients of variation even though they were no less likely than other segments with lower coefficients of variation to be due to the great coach effect. However, we still included it in the table because it allows for greater interpretability of the data. For instance, we can tell that Australia consistently performed well around 1996 to 2000 because of its relatively low coefficient of variation value of 0.182, even though it did not reach our requirements to extend the segment.

4.2 Estimating Impacts of Great Coaches

Our model from 4.1 gave us many valuable insights that can be used not only to find potential instances of the great coach effect but also to estimate the impacts of these great coaches. For each of the top 20 country-sport pairs we found with the highest z-scores, we calculated C:

$$C = (P_s - \mu \cdot L) \cdot 100\%.$$

This is the percentage of points that can be assumed to be due to the great coach effect out of the country's historical point count for that sport.

Table 17: Top 10 Countries with Highest Contributions in Olympic Sports

Year		Contribution (%)	Country (Sport)
Start	**End**		
1988	2012	100.0	Sweden (Handball)
1980	2004	100.0	United States (Synchronized Swimming)
1996	2016	88.9	Russia (Modern Pentathlon)
1984	2004	80.8	South Korea (Handball)
1968	1988	59.6	Hungary (Weightlifting)
1904	1908	55.0	Great Britain (Tug-Of-War)
1924	1928	50.0	Germany (Water Polo)
1908	1912	40.0	Netherlands (Fencing)
2004	2008	38.9	China (Trampolining)
1972	1976	33.3	Romania (Handball)

Based on this table, we can see that the dominance of the United States' synchronized swimming teams from 1984 to 2004 and of Sweden's handball teams from 1992 to 2012 accounted for 100% of their points over their entire history competing at the Games for those sports. Indeed, great coaches had a hand in these dominant runs: Gail Emery—whose athletes won 14 gold and silver medals in the Olympics and 23 gold and silver medals in World Championships[5]—coached the United States synchronized swimming team from 1988 to 1996, and Bengt Johansson coached the Swedish handball team from 1988 to 2004. In fact, Johansson's coaching ability was so influential that, under his leadership, the Swedish handball team were known as the Bengan Boys, after his first name.[6]

Another notable detail is that, as soon as these great coaches left their positions, their teams experienced a negative change point, immediately ending their segments of increased performance. This suggests that the influence of coaches may only last during their direct involvement and that great coaches have limited influence in improving their teams to the extent that they continue to improve even after the coaches have left.

4.3 Identifying Country-Sport Pairs Fit for Great Coaches and Estimating Impacts

Similarly, we can interpret the results of our model from 4.1 yet another way to identify country-sport pairs that would particularly benefit from investing in great coaches. For the same set of the top 20 country-sport pairs we found with the highest z-scores that we used in 4.2, we calculated ΔP_{norm}:

$$\Delta P_{\text{norm}} = \frac{C}{L} = \frac{P_s - \mu \cdot L}{L}.$$

This is the value that great coaches add to the given sport, normalized per Games so they can be compared with values for different sports. Countries should invest in coaches for sports where performances have been shown to be greatly affected by great coaches. It is also reasonable for countries to invest in great coaches for sports that they have previously excelled in under the leadership of great coaches in hopes of replicating previous success.

Table 18: Top 10 Countries with Highest ΔP_{norm} in Olympic Sports

Year		ΔP_{norm}	Country (Sport)
Start	**End**		
1980	1984	42.0	Romania (Rowing)
2000	2004	20.0	Russia (Weightlifting)
1996	2000	19.0	Russia (Diving)
2004	2008	14.0	China (Trampolining)
2000	2004	12.0	China (Taekwondo)
1904	1908	11.0	Great Britain (Tug-Of-War)
1968	1988	6.8	Hungary (Weightlifting)
1980	2004	6.7	United States (Synchronized Swimming)
2004	2008	6.0	Canada (Trampolining)
1992	1996	6.0	United States (Softball)

From this table, it is important to note that the 1984 Olympics were an anomaly because it took place in Los Angeles and was the target of boycotts from many of the Eastern Bloc nations, which played a major role in the anomalous success of the Romanian rowing teams.[7]

Aside from this outlier, Russia could benefit from investing in a great coach for weightlifting: in 2004, Russia's weightlifting athletes were coached by David Rigert, who set 65 world records and won an Olympic gold himself.[8]

China could benefit from investing in a great coach for trampolining, based on the success of its 2008 athletes.

Hungary could benefit from investing in a great coach for weightlifting, based on the success of its athletes from 1972 to 1988.

Outside of these three countries, it seems that weightlifting and trampolining are especially high return-on-interest investments for great coaches, as they both appeared twice in this list of the highest values added by great coaches.

Conveniently, these ΔP_{norm} are precisely the predicted points that these countries would gain if they invested in great coaches: we predict that Russia's investing in a great weightlifting coach could lead to up to 20 points, that China's investing in a great trampolining coach could lead to up to 14 points, and that Hungary's investing in a great weightlifting coach could lead to up to 6.8 points.

5 Problem 3: Insights into Olympic Medal Counts

Purely based on the regression results, it is clear that the United States dominating is sensible given their previous record and their host status in 2028. Some unexpected results include Germany and China's performance,

as China worsened in medal count quite a bit from 2024 and Germany came out of nowhere to take second place in gold medals. There are multitudes more observations like this to make, and they can be predicted to be due to any amount of reasons. Since the model accounts for unobserved effects, it's entirely possible that this is representative of some culture change or effect of China's governmental policy, or really anything under the sun that can be thought up.

Though we tried our best to predict the medal counts, there is a large confidence interval associated with the predictions, so in the end it is likely true that Number of Athletes, Host Status, and Past Performance are not sufficient indicators (at least through our method of regression) of Olympic medal performance. We suspect this is due to a lack of an economic/political factor, since that understandably has a major impact and there has been many cases where politics (such as political regimes) have affected the outcome of the Olympics much more than these predictors have. In addition, for developing countries (especially ones on track to earn their first medal), economic factors are many times more important than number of athletes or past performance, and countries at this level will definitely not be hosting. There is a large discrepancy in the number of countries performing poorly at the Olympics. An interesting question is whether this is a good thing or a bad thing: does the dominance of the United States and other developed countries motivate or demotivate developing countries to try to place in future Olympics? From a first world perspective, it may seem like we are giving them a chance to participate, and we are definitely doing so for the select few great talents from these countries. However, resources are overwhelmingly skewed for developed countries as are medal counts, as is especially highlighted in our 2028 predictions. Is it not in the interest of both high caliber players and coaches to gravitate towards where these resources are concentrated? The Olympic Committee should carefully consider how sports are chosen and how resources are allocated (to the extent they can), as this is very influential to the medal distribution.

6 Strengths and Weaknesses

6.1 Hierarchical Regression Model

A strength of our model is that, by implementing mixed effects and clusters, it is able to account for various levels of unobserved random effects that can influence medal counts. There are many factors that influence how a country performs at the individual competitor level or national level, so many that we cannot account for them all. So, our model implements random variables at the cluster and individual country level for each parameter to attempt to account for this, allowing for a more realistic prediction that isn't just purely data driven. In addition, our model appropriately accounts for the zero-inflation/censoring of Olympic medal data by using a Hurdle model approach with a probit regression to find probability of earning 0 medals or not.

One weakness about this model is it's fit. Maximizing the Log Likelihood function would lead to a good fit, but during implementation the optimization did not converge for many methods of optimization. This inevitably led to log likelihoods and AIC/BIC fits that were subpar. Though we tried to account for this by weighting one of the clusters to reduce heteroscedasticity, it didn't help the convergence issue. This may imply the nature of the data is different from how our model assumes it to be (in fact it is, since we tested the data to still be heteroscedastic even after improvements). One improvement we explored was to use a **Hamiltonian Monte Carlo** optimization to replace MLE, but this approach did not end up working. Some approach that accounts for the heteroscedasticity error would improve the model significantly.

6.2 Time Series Analysis with Moving Average Model

Our model's main strength lies in its moving average, which allowed us to interpret anomalous performances in the context of performances immediately before and after it. This, in turn, allowed us to create segments of increased performance, which we interpreted as possible signs of the great coach effect. This was much

more effective than identifying individual anomalous performances by considering only the population mean and standard deviation instead of the rolling statistics we used.

The main weakness of our model is that there is no way to determine whether anomalous performances were due to the great coach effect or any number of outside influences without manually comparing our findings to historical great coaches. Because of the limited data, this is inevitable.

7 Conclusion

The Olympics is not only a stage for athletes to show off their skills, but for countries to establish their presence in the world. Developing countries are increasingly entering the Olympics, and their participation adds a new layer of competition and uncertainty to the Games. As the Olympics continue to grow in scope and diversity, the ability to predict medal counts becomes increasingly complex. By leveraging statistical models to investigate data on Olympic medal distributions, we have explored the nature of such data and how it is affected by great coaches.

8 Works Cited

1. Swaddling, J., & British Museum. Trustees. (2008). *The ancient Olympic games (2nd ed., rev.updated.)*. University of Texas Press.

2. Scelles, N., Andreff, W., Bonnal, L., Andreff, M. and Favard, P. (2020), Forecasting National Medal Totals at the Summer Olympic Games Reconsidered. *Social Science Quarterly*, 101: 697-711. https://doi.org/10.1111/ssqu.12782

3. Costeiu, A. (2012, May 30). *Voina's Last Two Games for Romania.* European Handball Federation. https://www.eurohandball.com/en/news/en/voina-s-last-two-games-for-romania/

4. *Patience Is the Key to Talent Identification and Development: Olympic Trampoline Coach*. (2021, October 14). Sport for Life. https://sportforlife.ca/blog/patience-is-the-key-to-talent-identification-and-development-olympic-trampoline-coach/

5. *Gail Emery.* (2024, March 25). International Swimming Hall of Fame. https://ishof.org/honoree/honoree-gail-emery/

6. Bruun, P. (2013, September 9). *Memories of a "Bengan Boy."* European Handball Federation. https://www.eurohandball.com/en/news/en/memories-of-a-bengan-boy/

7. Kobierecki, Micha l. (2015). Boycott of the Los Angeles 1984 Olympic Games as an Example of Political Play–Acting of the Cold War Superpowers. *Polish Political Science Yearbook*. 44. 10.15804/ppsy2015008.

8. *David Rigert.* (n.d.). Lift Up. Retrieved January 27, 2025, from http://www.chidlovski.net/liftup/l_galleryResult.asp?a_id=217