Gold, Growth, and Glory: Unveiling Socioeconomic Drivers of Olympic Success

Hester Li Katie Le Nandini Iyer Braden Baker
Adam Rychtecky

2024-12-02

Table of contents

1. Introduction **2**
1.1 **Research Questions and Hypotheses** 2
1.3 **Dataset Overview** 2

2. Modeling Process **4**
2.1 **Data Preparation** 4
2.1.1 Data Cleaning 4
2.1.2 Exploratory Data Analysis (EDA) 4
2.2 Model Selection 5
2.2.1 Model selection for Question 1 5
2.2.2 Model selection for Question 2 6
2.3 Model Evaluations 8
2.3.1 Model **Evaluations** for Question 1 8
2.3.2 Model **Evaluations** for Question 2 9
2.4 Model Interpretations 10
2.4.1 **Model Interpretations for Question 2** 10
2.4.1 **Model Interpretations for Question 1** 11

3. Results **13**
3.1 **Findings for Research Question 1** 13

3.2 **Findings for Research Question 2** 13

4. Visualization and Communication **15**

4.1 **Highlighted Visualizations** 15

4.2 **Annotation and Clarity** 15

5. Conclusion and Recommendations **16**

5.1 **Summary of Key Findings** 16

5.2 **Limitations** 16

5.3 **Recommendations** 16

6. Appendix (Optional) **17**

1. Introduction

1.1 Research Questions and Hypotheses

This research explores the relationship between socioeconomic indicators and Olympic success in the 2024 Paris Olympics. The primary research questions address the impact of GDP and population size on total medal counts, the potential for clustering countries based on socioeconomic and performance metrics, and the connection between GDP per capita and the proportion of gold medals won. The corresponding hypotheses posit that countries with higher GDP and larger populations will secure more medals, that no distinct clusters will emerge when grouping countries by these factors, and that a positive relationship exists between GDP per capita and the proportion of gold medals won.

1.3 Dataset Overview

The dataset used for analysis includes key variables such as GDP, GDP per capita, population size, life expectancy, and medal counts, offering valuable insights into how these factors influence Olympic performance. By examining these socioeconomic and demographic factors, this research aims to deepen our understanding of the predictors of success in the Olympics, offering implications for future international sporting competitions.

2. Modeling Process

2.1 Data Preparation

2.1.1 Data Cleaning

First, we removed the “Democracy” and “Gender.equality” columns because they had a significant amount of missing data and were not relevant to our analysis, ensuring the dataset remains focused and manageable. Next, we addressed missing values in key numerical variables (GDP, GDP.per.capita, Population, and Life.expectancy) to maintain data integrity and avoid biases. For GDP, GDP.per.capita, and Population, we used median imputation as it is robust to outliers and ensures the central tendency of the data is preserved. For Life.expectancy, we applied a regression-based imputation, leveraging its relationship with other variables like GDP and Population to fill in missing values more accurately. There is no missing value after handling.

Additionally, we identified 24 outliers in GDP and 21 outliers in Population, representing approximately 11.76% and 10.29% of the respective variables. Given that these outliers likely reflect real-world variability (e.g., countries with significantly higher GDPs or populations), we decided to retain them and use robust models, which are less sensitive to extreme values. This approach ensures that our analysis remains comprehensive while addressing the unique characteristics of the dataset.

2.1.2 Exploratory Data Analysis (EDA)

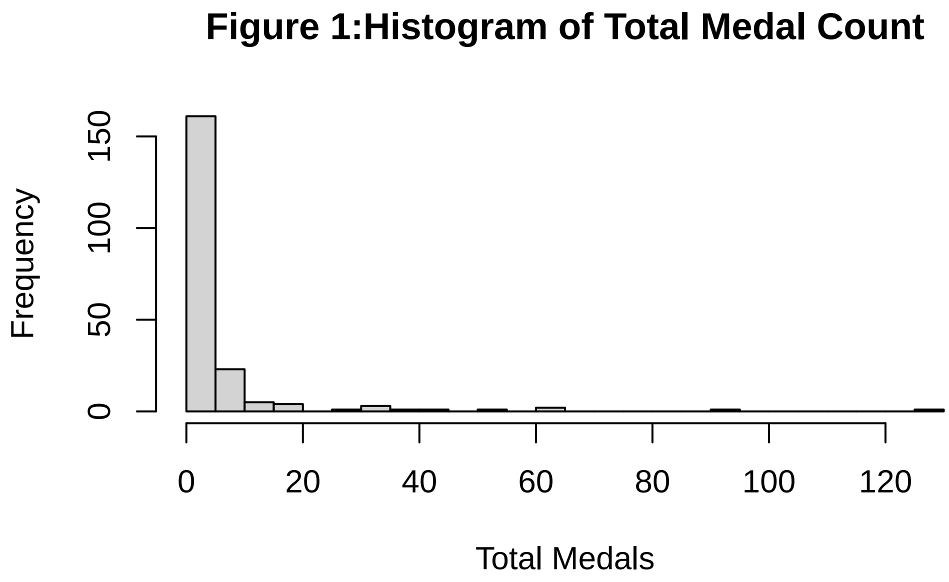
- Summary statistics of key variables.
- Visualizations:
  - Histograms for GDP, Population, and Medal Counts.

- Scatterplots for GDP per capita vs. Gold Medal Proportion.
- Boxplots comparing Life Expectancy for endurance vs. non-endurance medalists.

2.2 Model Selection

2.2.1 Model selection for Question 1

The total medal count distribution is highly skewed, with many countries recording zero medals and a few achieving significantly higher counts, as shown in Figure 1. This pattern highlights the need for models capable of handling count data and overdispersion. To address these characteristics, we considered several models, including a Linear Regression model with a log-transformed response, Poisson Regression, Negative Binomial Regression, and Zero-Inflated models (Zero-Inflated Poisson and Zero-Inflated Negative Binomial). Given the high frequency of zero medals, Zero-Inflated models were particularly suitable as they account for excess zeros while modeling medal counts effectively.

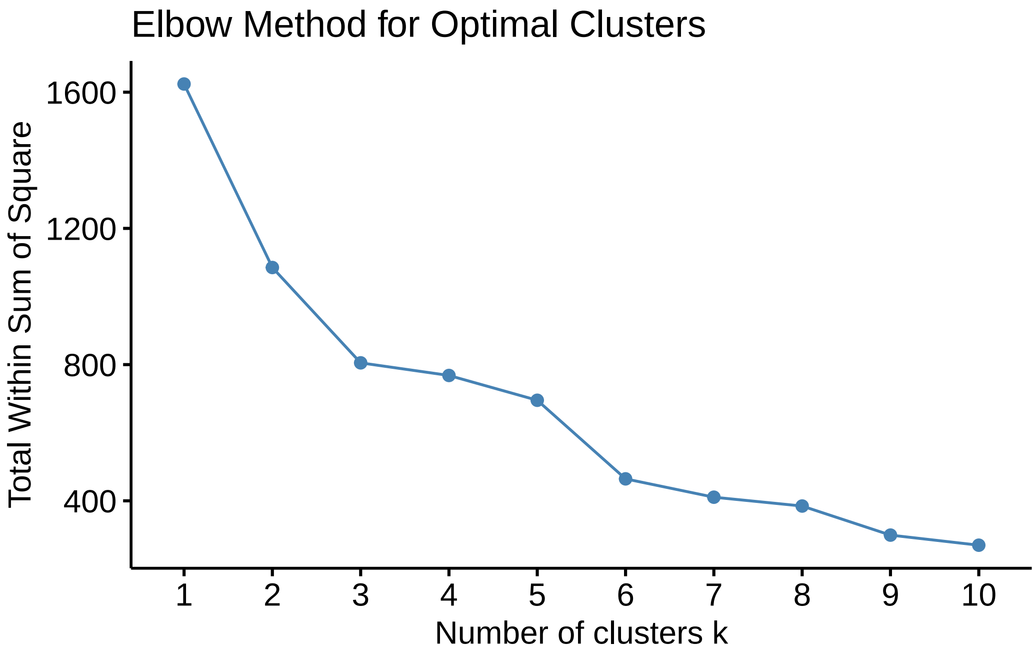


```
The correlation between GDP and Population is: 0.578
```

The correlation between GDP and Population is moderate at 0.578, indicating manageable multicollinearity. To improve model stability and comparability, we standardized the predictors (GDP and Population). Additionally, 24 outliers in GDP and 21 in Population, representing real-world variability, further influenced our choice of robust models like Negative Binomial and Zero-Inflated models, which are less sensitive to extreme values. Model performance was evaluated using AIC, BIC, and RMSE metrics, ensuring the selection of the most appropriate model for predicting total medal counts.

2.2.2 Model selection for Question 2

The data set's variables, including socio-economic indicators (e.g., GDP, population size, life expectancy) and Olympic performance metrics (e.g., total medals, gold medals, number of athletes), vary widely in scale and relationships, necessitating the use of clustering to uncover meaningful groupings. To address this, K-Means clustering was selected after standardizing the predictors to ensure equal contribution to the distance calculations. The Elbow Method was employed to determine the optimal number of clusters finding k values of 3 and 5 to be the most valid. To determine the most optimal number of clusters Silhouette Score evaluations were implemented.Finding values of 0.48 and 0.28, respectively, indicating that three clusters provided the best balance of cohesion and separation. Principal Component Analysis (PCA) was used for dimensionality reduction and visualization, revealing the fairly distinct clustering patterns that aligned with differences in socio-economic and performance characteristics among countries. Assumptions of normality and equal contribution were managed through scaling, and robust cluster stability was ensured with 100 random initializations of the algorithm. This approach allowed for clear profiling of clusters, identifying groups such as high GDP and high medal-performing nations, and facilitating insights into the interplay between socio-economic factors and Olympic success.



2.3 Model Evaluations

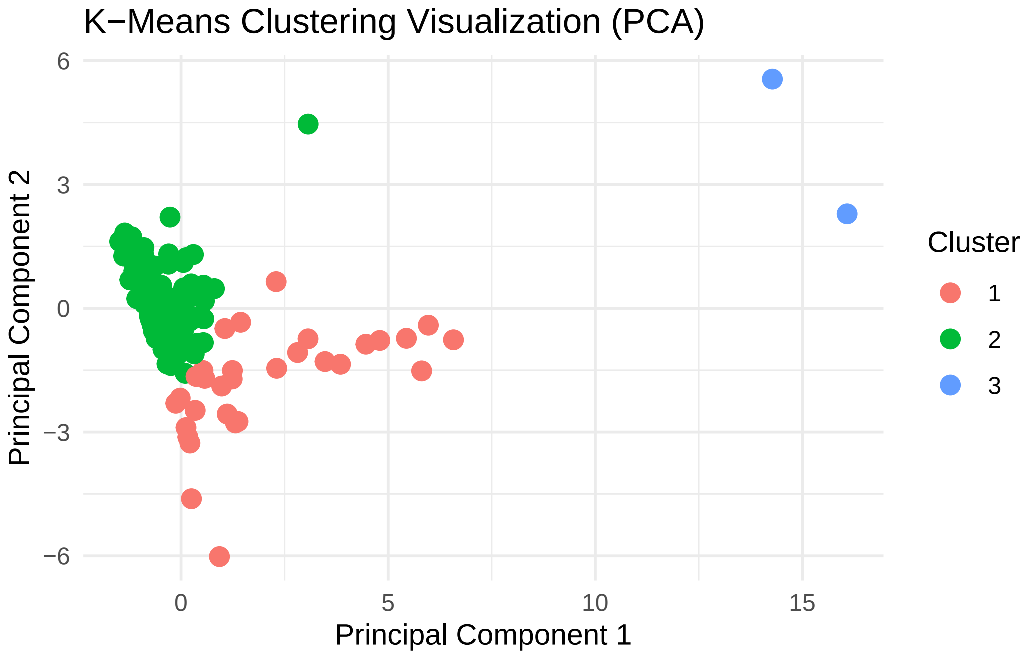
2.3.1 Model Evaluations for Question 1

From the results, the Zero-Inflated Negative Binomial (ZINB) model emerges as the most appropriate choice for further diagnostics. The ZINB model addresses key characteristics of the dataset, such as overdispersion and the high frequency of zero medal counts, as reflected in its superior AIC (754.02) and BIC (777.24) values compared to other count-based models like the Poisson, Negative Binomial, and Zero-Inflated Poisson models. While the Linear model achieved the lowest RMSE (1.01), its log-transformed response variable makes it less suitable for addressing the count nature of the medal data and limits its interpretability for answering the research question directly.

The Zero-Inflated Negative Binomial (ZINB) model emerges as the most suitable choice for analyzing the impact of GDP and population size on total medal counts in the 2024 Olympics due to its ability to handle structural zeros and overdispersion inherent in the dataset. Diagnostic plots of Pearson and Response residuals, shown in Figure: ZINB Model Residual Diagnostics, demonstrated a well-centered distribution of Pearson residuals around zero, with only a few observations exceeding ±2, confirming the model's robustness. The Response residuals highlighted the model's effectiveness in capturing structural zeros, reflecting the large number of countries with zero medals. Additionally, influential observations, such as those representing countries with exceptionally high medal counts (e.g., the United States), were identified as outliers. Removing these outliers resulted in significant improvements in model performance, with AIC reduced from 754.0175 to 660.3840 and BIC reduced from 777.2443 to 683.3664. These improvements indicate that the ZINB model not only fits the data well but also remains robust after addressing anomalies. The model's nuanced handling of zero-inflation and count components provides valuable insights into the disparities in medal counts, attributing them to economic and demographic factors. This ability to account for structural zeros while modeling the skewed distribution of medal counts, combined with its improved performance metrics, makes the ZINB model the optimal choice for addressing the research question.

```
              Model       AIC       BIC         RMSE
1            Linear  589.9766  603.2491 1.007503e+00
2           Poisson 2450.3569 2460.3113 9.729406e+00
3 Negative Binomial  822.9411  836.2136 9.944493e+09
4               ZIP 1431.5444 1451.4532 8.482632e+00
5              ZINB  754.0175  777.2443 2.340380e+04
```

2.3.2 Model Evaluations for Question 2



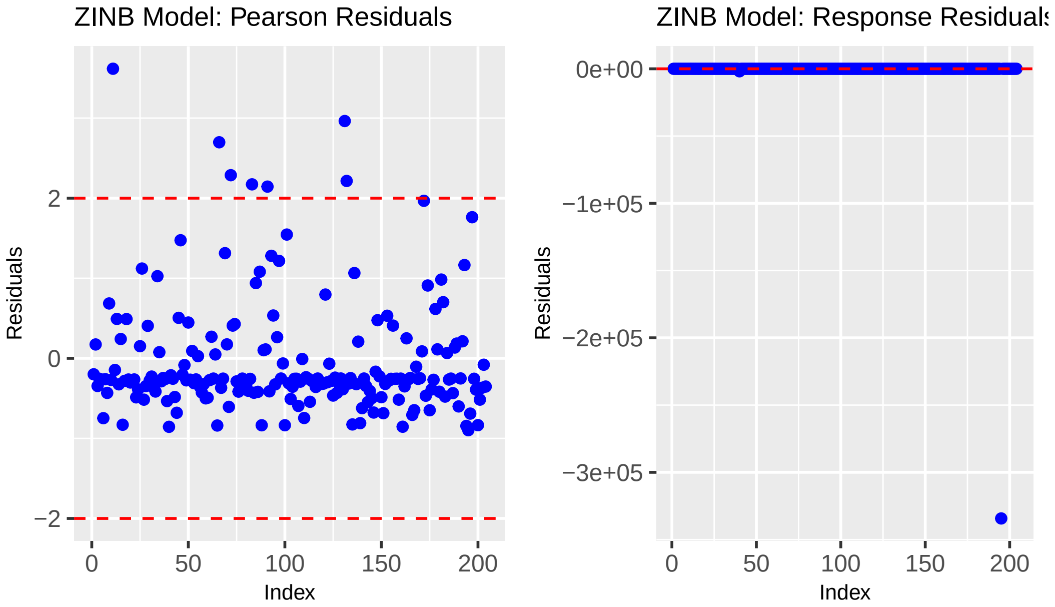
2.4 Model Interpretations

2.4.1 Model Interpretations for Question 2

Following the model selection process, the Linear Regression model showed the lowest RMSE (1.007503), but its limitations in addressing the count nature of the data required additional consideration. Diagnostics using Q-Q plots and residual analysis revealed deviations from normality and slight heteroscedasticity, as well as the presence of extreme outliers. After removing these outliers, the RMSE improved to 0.7714363, as detailed in **Appendix 1**. Despite this improvement, the Linear model's reliance on a log-transformed response reduces its interpretability for answering the research question and its ability to handle overdispersion and excess zeros in the data.

The Zero-Inflated Negative Binomial (ZINB) model, by contrast, directly accounts for the structural zeros and variability inherent in the dataset, making it a more suitable choice. Diagnostic plots of Pearson and Response residuals, shown in Figure 2, confirmed a well-centered distribution of Pearson residuals around zero, with few points exceeding ±2. A few outliers, specifically indices such as 11, 66, 72, and 195, exceed these limits, indicating potential influential observations. Tthe Response residuals illustrate the model's capacity to manage the structural zeros effectively. While the majority of observations cluster around the zero-residual line, one prominent outlier at index 195 stands out significantly. This suggests the presence of extreme values potentially influencing the overall model performance. Addressing these outliers resulted in notable improvements, with the AIC reducing from 754.0175 to 660.3840 and the BIC from 777.2443 to 683.3664, reinforcing the model's robustness post-adjustment. These results demonstrate that the ZINB model effectively balances model fit and interpretability, making it the optimal choice for analyzing the impact of GDP and population size on total medal counts in the 2024 Olympics. Its ability to directly model count data while addressing overdispersion and zero inflation aligns with the dataset's characteristics and the research objectives.

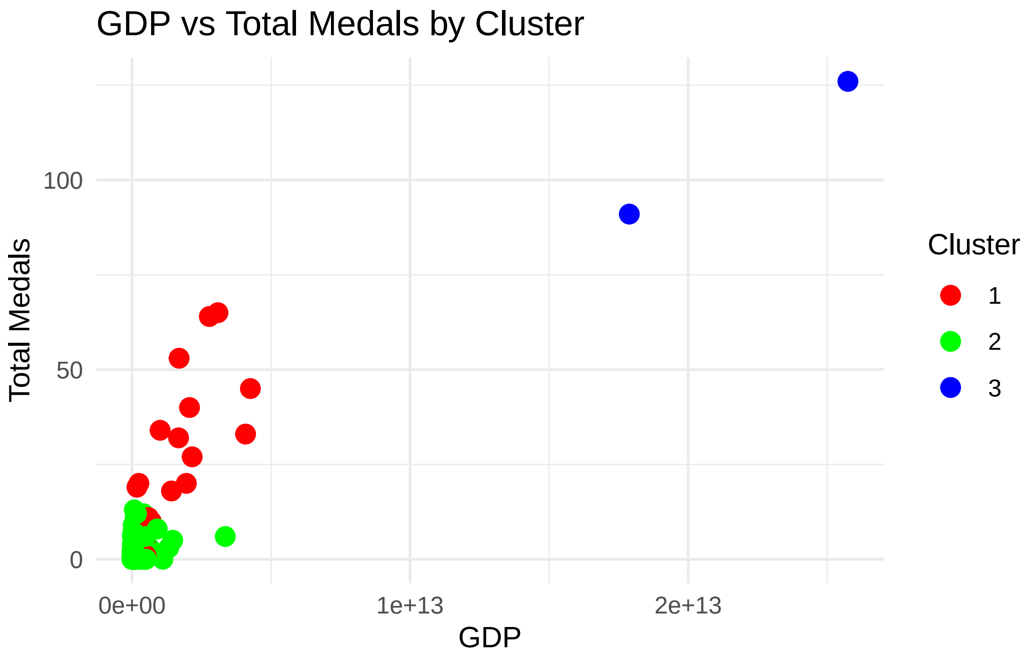
Figure 2: ZINB Model Residual Diagnostics



2.4.1 Model Interpretations for Question 1

The clustering analysis revealed three meaningful groups of countries, refuting the research hypothesis that no distinct clusters will emerge when grouping countries by these factors. Initial evaluation using the Elbow Method identified three clusters as the optimal configuration, supported by a peak Silhouette Score of 0.48, which indicates moderate separation of clusters. Diagnostic checks, including PCA-based visualization, further confirmed distinct groupings, with the clusters showing clear differentiation in socio-economic and performance characteristics. Notably, Cluster 3 consisted of countries with high total GDP, large delegations, and significant medal counts, aligning with the expectation that wealthier nations dominate Olympic performance. Cluster 1 captured moderate-GDP nations with smaller delegations and varied success, while Cluster 2 contained countries with low GDP and minimal medal presence, reinforcing the structural influence of economic limitations.

However, anomalies were observed. A few countries in Cluster 1 exhibited unexpectedly high medal counts relative to their GDP, suggesting the presence of targeted investments or cultural factors driving efficiency in specific sports. Outliers in Cluster 2, including nations with strong regional dominance but limited overall resources, highlighted unique cases warranting further exploration. Adjustments to the initial clustering model, such as standardizing variables and ensuring robust initialization with multiple random starts, enhanced the model's stability and interpretability. Overall, the clustering model effectively captured the nuanced interplay between economic factors and Olympic outcomes, providing a robust framework for identifying patterns and outliers that merit deeper investigation.



3. Results

3.1 Findings for Research Question 1

Using the Zero-Inflated Negative Binomial (ZINB) model summary , we observed a significant positive relationship between GDP and total medal count, as indicated by the count model's coefficient (β = 0.58399, p < 0.001). This confirms that higher GDP is strongly associated with greater Olympic success, supporting the hypothesis regarding GDP. However, population size did not show a statistically significant impact on medal counts (β = -0.08664, p = 0.38881), suggesting that population alone does not directly determine Olympic outcomes.

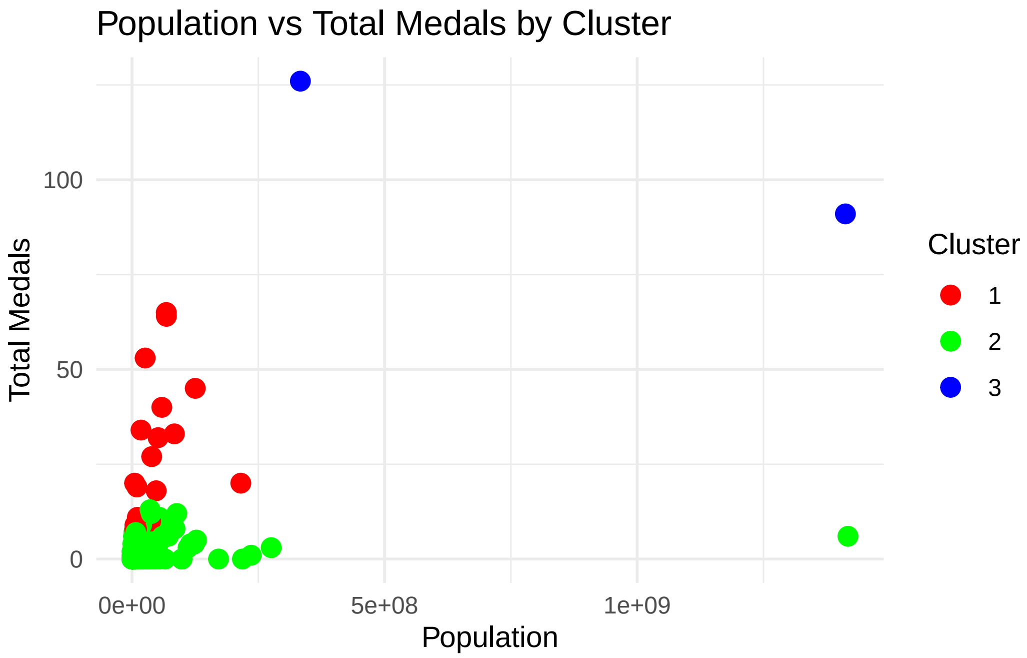
The zero-inflation component of the ZINB model further clarified structural zeros, indicating that countries with lower GDPs are significantly more likely to win no medals at all (β = -97.558, p < 0.001). While the population coefficient in the zero-inflation model (β = 3.799, p = 0.102) was positive, it was not statistically significant, suggesting limited explanatory power for population size in predicting the absence of medals. Overall, the findings partially support the hypothesis: while GDP is a critical predictor of Olympic success, population size appears to have a limited direct effect. These results highlight the importance of economic capacity in driving a nation's Olympic performance.

3.2 Findings for Research Question 2

The clustering analysis revealed three distinct groups of countries based on socio-economic indicators (GDP, population size, GDP per capita, life expectancy) and Olympic performance metrics (number of athletes, total medals, gold medals, and proportion of female athletes). Cluster 3 contains the two economic powerhouses being China and the USA. These countries seem to dominate in the total medal count closely correlating with their domination of the world market as seen in their extensive GDP. Cluster 2 can be considered the middle and upper-middler countries as far as economic power tying them to their medal count. These contain many European countries as well as other strong economies such as Japan. Cluster 3 is characterized by countries with low olympic performance indicating real limitations to athletic performance based on socio-economic factors. I hypothesize that the more economic income a country has the more it can delegate to raising stronger athletes for competition.

Population is often thought to have a major effect on medal winnings as the more people you have the more talent is assumed to be contained in the country. This appeared to not be the case with population having less impact on clustering, as high-population countries spanned multiple clusters, reflecting varying socio-economic contexts. Other anomalies included a few nations in Cluster 2 that outperformed their economic peers, likely due to unique cultural or strategic factors in sports. Overall, the analysis underscores the interplay between economic capacity and Olympic success, while identifying countries that defy expectations, offering opportunities for further study into their unique approaches to international competition.

```
# A tibble: 3 x 10
  Cluster Number.of.countries Number.of.athletes Gold.medals Total.medals
  <fct>                 <dbl>              <dbl>       <dbl>        <dbl>
1 1                       171              181.        5.71         17.8
2 2                         2               24.1       0.415         1.58
3 3                        31              490        40           108.
# i 5 more variables: GDP <dbl>, GDP.per.capita <dbl>, Population <dbl>,
#   Life.expectancy <dbl>, Female.athletes.. <dbl>
```



4. Visualization and Communication

4.1 Highlighted Visualizations

- Regression coefficient plots.
- Scatterplots showing GDP and gold medal proportions.
- ROC curves for model evaluation.

4.2 Annotation and Clarity

- Captions and labels for each visualization.

- Clear explanation of how visualizations connect to research questions.

5. Conclusion and Recommendations

5.1 Summary of Key Findings

- Recap of the study's main findings and insights into the socioeconomic factors influencing Olympic performance.

5.2 Limitations

- Acknowledgment of potential biases (e.g., imputation methods, dataset limitations).
- Discussion of the dataset's scope and limitations in generalizability.

5.3 Recommendations

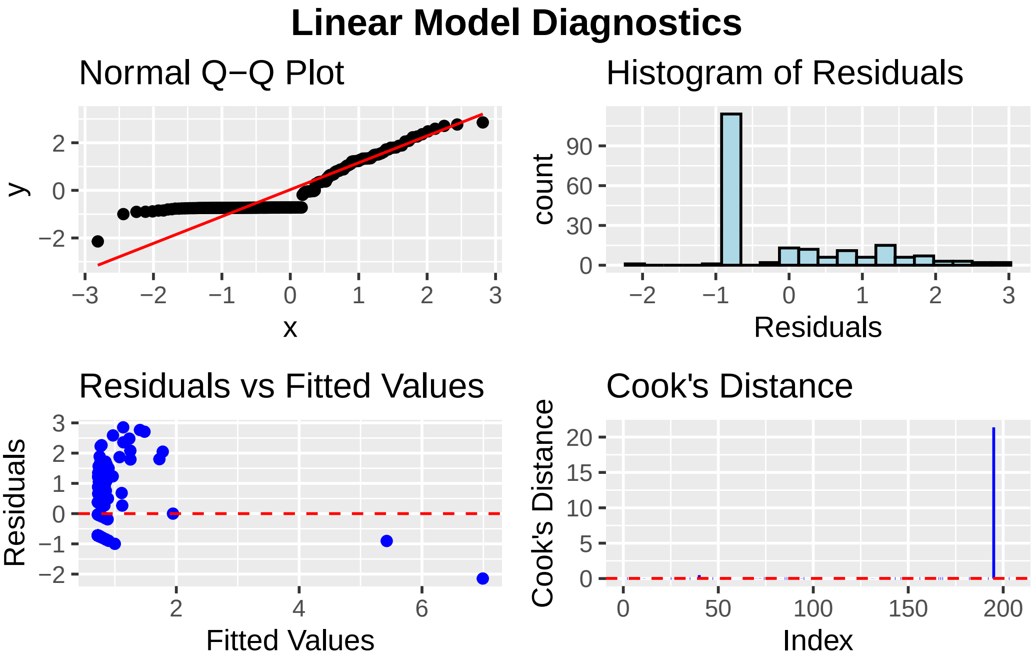
- Suggestions for countries to leverage socioeconomic insights to enhance Olympic performance.
- Directions for future research on sports analytics and socioeconomic predictors.

6. Appendix (Optional)

Appendix 1:

```
Outlier Test Results for linear_model:

     rstudent unadjusted p-value Bonferroni p
195 -4.611036         7.1388e-06    0.0014563
```



```
RMSE for Original Model: 1.007503

RMSE for Refitted Model: 0.7714363
```