JSC Midterm Report

1. Introduction

1.1 Introduction

The Olympics is one of theworld's greatest sporting events that occurs once every 4 years, where athletes from across the globe compete for national pride and athletic excellence. While factors such as training and talent plays a significant role in an athlete's success, economic conditions may also influence a country's ability to produce medal-winning athletes. Wealthier countries may have better sports facilities, higher government investments in athletics, and greater access to elite coaching, which could contribute to higher Olympic medal counts. Thus, in this project, I would like to explore the relationship between economic indicators and medal counts in a country, aiming to identify which economic factors best predict the number of medals a country wins.

To investigate this, I will use two datasets:

1. Historical Olympic Medals Dataset

The Kaggle dataset,"Historical Olympic Medals Data (1994-2024)", provides medal counts for each nation across multiple Olympic Games. This dataset includes information on the number of gold, silver, and bronze medals won by each country.

2. WorldBank Open Data API

This API is used to retrieve economic indicators such as GDP per capita, education expenditure, health expenditure, population size, and unemployment rate for each country per year. These indicators provide insight into a country's overall economic strength, investment in human capital, and potential capacity to support athletic programs.

I will merge these two datasets to create one data frame that shows the total number of medals won and the economic performance of a country in a specific year.

So the research question that I will answer in this project is: "How do economic conditions influence a country's success in the Olympics, and which indicators best predict medal performance". This question will allow us to assess how different economic factors correlate with a country's performance, providing insights into the role of economic success in global sports competition. Additionally, I am planning to add a predictive modeling section, using the 2024 Olympic data to do the following:

1. Model Development: Multiple models will be created to predict medal count based on a country's economic measures, using data from 2000-2022.
2. 2024 Medal Predictions: Using 2024's economic performance, the developed models will generate predictions for the number of medals each country is expected to win.
3. Model Evaluation: The predicted medal counts will be compared to the actual 2024 Olympic results to assess model accuracy and identify which economic factor best predicts a country's Olympic success. This will provide insights on how well economic success alone can predict a country's success in the olympics and explore which economic indicators are the strongest predictors of medal success.

My hypothesis for the first part of the question: "How do economic conditions influence a country's success in the Olympics" is that I expect country with high economic measures to perform well in the Olympics due to the investments and quality of training that the athletes can gain.

2. Introduction

2.1 Data Acquiring

As mentioned in the introduction I used WorldBank Open Data API for economic indicators and Historical Olympic Medals Dataset for medal counts.

For the WorldBank Open Data API, I had to make API calls to gather the data. The API call had a limit of 500 so I had to make multiple GET requests to retrieve a complete dataset. Furthermore, the API was queried separately for each economic indicator (GDP per capita, Government expenditure on education, Health expenditure, Total population, Unemployment rate). Each dataset contained the CountryCode in ISO3 format, the Year, and the respective economic indicator for that country in the specified year. Then I merged each data by CountryCode and year.

The Olympic Medal Data, was a kaggle dataset so I downloaded it and read it as a datatable in R. The datasets contained CountryCode in NOC, the number of Gold, Silver, and Bronze medals won, and the total number of medals won by that country. Since there was one csv file per year, I merged the datasets by Country and Year. One note is that, the dataset only contained entries for countries that won at least one medal for that year. Since countries that did not win medals were absent from the dataset, I padded them with 0 medals to ensure accurate analysis.

After preparing both datasets, I merged them by matching the CountryCode from the World Bank dataset with the NOC codes from the Olympic dataset. However, NOC codes (CountryCode used in olympics) differ from ISO3 country codes, so for those countries whose codes did not map, I had to manually map the NOC code to a ISO3 code. (Could not do this for all countries as there were quite a number of them). Then I merged the two datasets together by Year and CountryCode. Lastly, I downloaded a dataset that maps ISO3 country code to Country Names using the WorldBank Open Data API and merged them to add a CountryName column for readability.

2.2 Data Cleaning & Wrangling

After acquiring the dataset, several preprocessing steps were performed to ensure the data was clean, properly formatted, and is ready for analysis.

2.2.1 Enhancing Readability and Removing Unnecessary Columns To improve the interpretability of the dataset, the following changes were made:

- Renaming Columns:

Economic indicators obtained from the World Bank API were originally stored under their respective API codes. These were renamed for clarity. For example, "SE.XPD.TOTL.GD.ZS" was renamed to "Education_Expenditure"

- Adjusting the Population Scale:

Population values were originally recorded as raw counts which are too large to read and interpret efficiently. Thus, these values were converted into millions.

- Dropping Unnecessary Columns:

The original Olympic dataset contained separate columns for gold, silver, and bronze medals. Since this analysis only considers total medal count, these columns were removed. Table 1 below presents the final list of variables in the cleaned dataset that has 9 columns and 892 rows:

2.2.2 Handling NA observations

Our dataset initially contained a total of 357 NA values over 4 economic indicators: 35 in GDP per capita, 198 in Education Expenditure, 112 in Health Expenditure, and 12 in Unemployment Rate. However, removing them from the dataset would have significantly reduced the dataset, leading to a loss of valuable data. Therefore, instead of dropping all NA observations, I applied imputation to keep as much information as possible.

Table 1: Summary of Variables in the Dataset

Variables	Type	Description
CountryCode	character	ISO3 country code.
Year	numeric	Olympic event year.
Total_Medals	integer	Total medals won by the country in that year's Olympics.
GDP_per_capita	numeric	Gross Domestic Product per capita in constant 2015 USD.
Education_Expenditure	numeric	Government expenditure on education as a percentage of GDP.
Health_Expenditure	numeric	Total health expenditure as a percentage of GDP.
Population	numeric	Total population of the country in that year (in millions).
Unemployment_Rate	numeric	Unemployment rate as a percentage of total labor force.
CountryName	character	Full country name.

Imputation is the process of replacing missing values with estimated values based on the existing data. In this case, I used median imputation, where missing values for each economic indicator were replaced with the median value of that country. This approach is effective because the median is resistant to outliers, unlike the mean and it provides a reasonable estimate without introducing a huge artificial bias as values are not generated.

However, this approach is problematic when a country is missing a significant portion of its data for an economic indicator, as imputing in this scenario would introduce too much uncertainty. To address this, I decided to remove countries that had had more than 6 NA values in at least one economic indicator. I decided to use 6 as the threshold because there were 12 Olympic Games from 2000 to 2022, so if a country has more than 6 NA values, it means the country has more than half of its data missing for an economic indicator. This threshold ensures that only countries with a reasonable amount of data for imputation remain in the dataset.

This approach led me to remove a total of 10 countries from the dataset, such as North Korea, Colombia, and Channel Islands. And after removing the NA values I decided to perform imputation to minimize the loss of data.

2.2.3 Identifying Problematic Observations

To check for any problematic observations, I decided to look at the summary statistics for the economic indicator variables. As the economic indicator variables were floating point numbers, the summary presented the minimum, maximum, 1st & 3rd quadrant, median, and mean. Most of the values presented were in a reasonable range without any negative values. However, there were two numbers that struck me: maximum GDP per capita of 99,677.47 and maximum unemployment rate of 26.71 as these values were much larger than the 3rd quartile presented in the summary.

To take a deeper look at these values and any other extreme observations I decided to filter the following observations: entries with GDP per capita greater than 90,000 and unemployment rate greater than 20%. Filtering for GDP per capita $> 90{,}000$ showed that there was only one row of extreme observation: Ireland in 2022 with GDP per Capita of 99677.47. After checking with external sources, I confirmed that this value accurately reflects Ireland's GDP. As a result, I decided to keep this observation.

Similarly, I decided to look at all of the data entries with an unemployment rate greater than 20%. This time there were a total of 18 rows, including countries such as Gabon, Greece, and Namibia to have an unemployment rate over 20% in certain years. After checking external sources, such as macrotrends.net, I came to the conclusion that these unemployment rates were legitimate values that occurred during economic crisis, so I decided to keep them in the dataset.

While reviewing the dataset, I also noticed that the maximum population of 1,425.42 million is extremely high. However, given that China and India both have populations exceeding 1.4 billion, this value is realistic.

Thus, while there were extreme points in the dataset, no data points were removed.

Table 2: Summary Statistics of Numeric Variables

Variable	Min	1st Quartile	Median	3rd Quartile	Mean	Max	# of NAs
Total_Medals	0.00	0.00	1.00	7.00	5.99	100.00	0
GDP_per_capita	253.69	2,923.33	8,567.91	31,419.75	17,153.68	99,677.47	0
Education_Expenditure	0.35	3.75	4.78	5.59	4.80	14.06	0
Health_Expenditure	1.82	4.71	6.70	8.78	6.79	15.53	0
Population	0.28	4.58	11.03	36.95	68.29	1,425.42	0
Unemployment_Rate	0.11	3.61	5.73	8.50	6.87	26.71	0

2.3 Data Cleaning & Wrangling

After cleaning and wrangling, summary statistics were computed for the updated dataset which contained 772 rows with 9 columns. The table below presents the minimum, first quartile (Q1), median, third quartile (Q3), mean, maximum, and number of missing values.

Preliminary Results

3.1 Top Performing Nations in Olympics

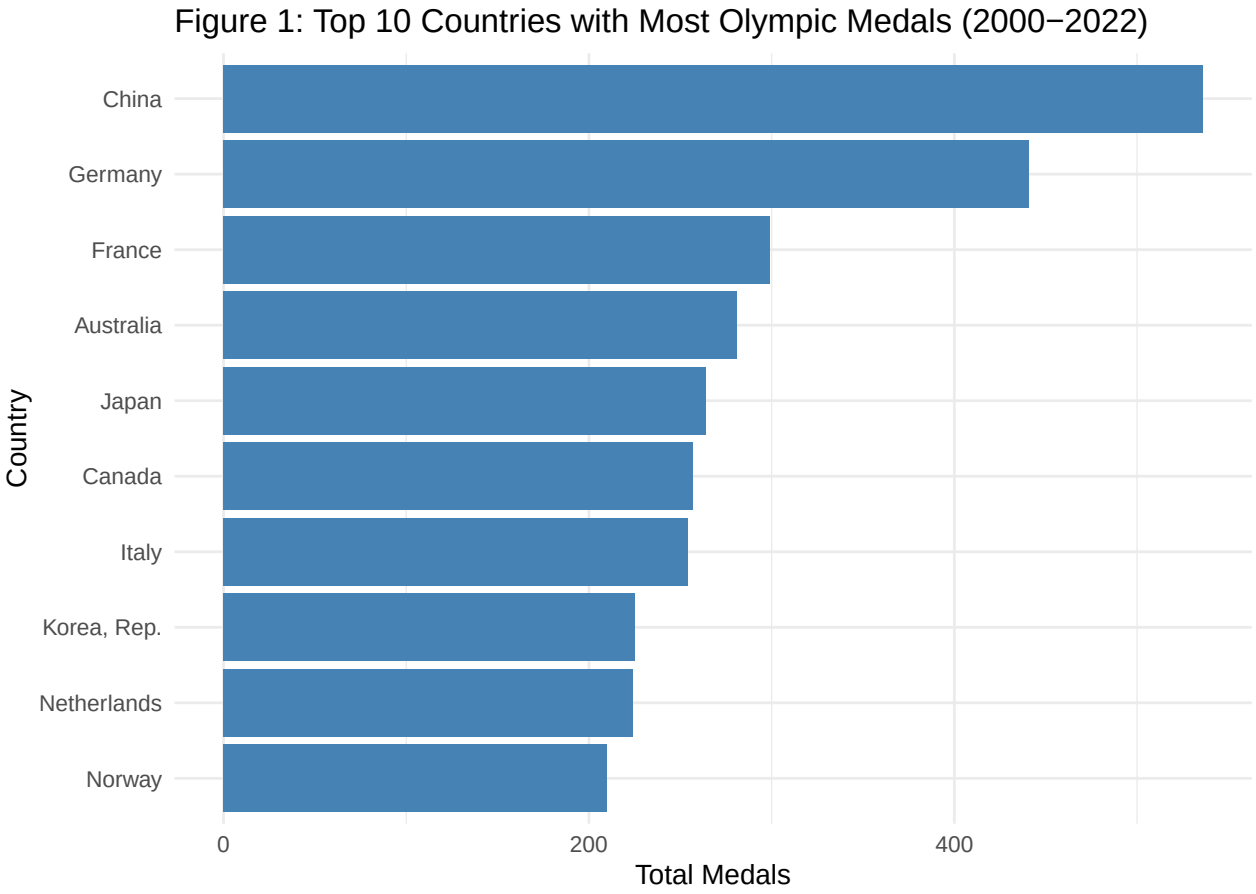

Figure 1 shows the top 10 countries with the highest total medal counts from 2000 to 2022. We see that China and Germany have a much higher highest medal count compared to the other countries, suggesting that these two nations have consistently ranked among the highest in total medals. Also, since Germany and China are considered to be "wealthy" countries with good economic performance, this result suggests that

having a good economic performance could affect Olympic success.

It is worthy to note that countries such as France, Australia, and Japan also rank among the top medal winners, despite having smaller populations than countries like the United States or Russia, which are missing from this ranking. This suggests that factors beyond total population affect Olympic Success, hinting at the possibility that economic performance could be playing a factor here.

We also see that Norway, a relatively small country in terms of population, ranks in the top 10. This could be because Norway is a country that is extremely strong in Winter Sports. This reinforces the idea that while economic strength is important, other cultural and environmental factors also play a role in shaping a country's Olympic success.

3.2 Relationship Between GDP Per Capita and Olympic Medals

Figure 2: Relationship Between GDP Per Capita and Olympic Medals

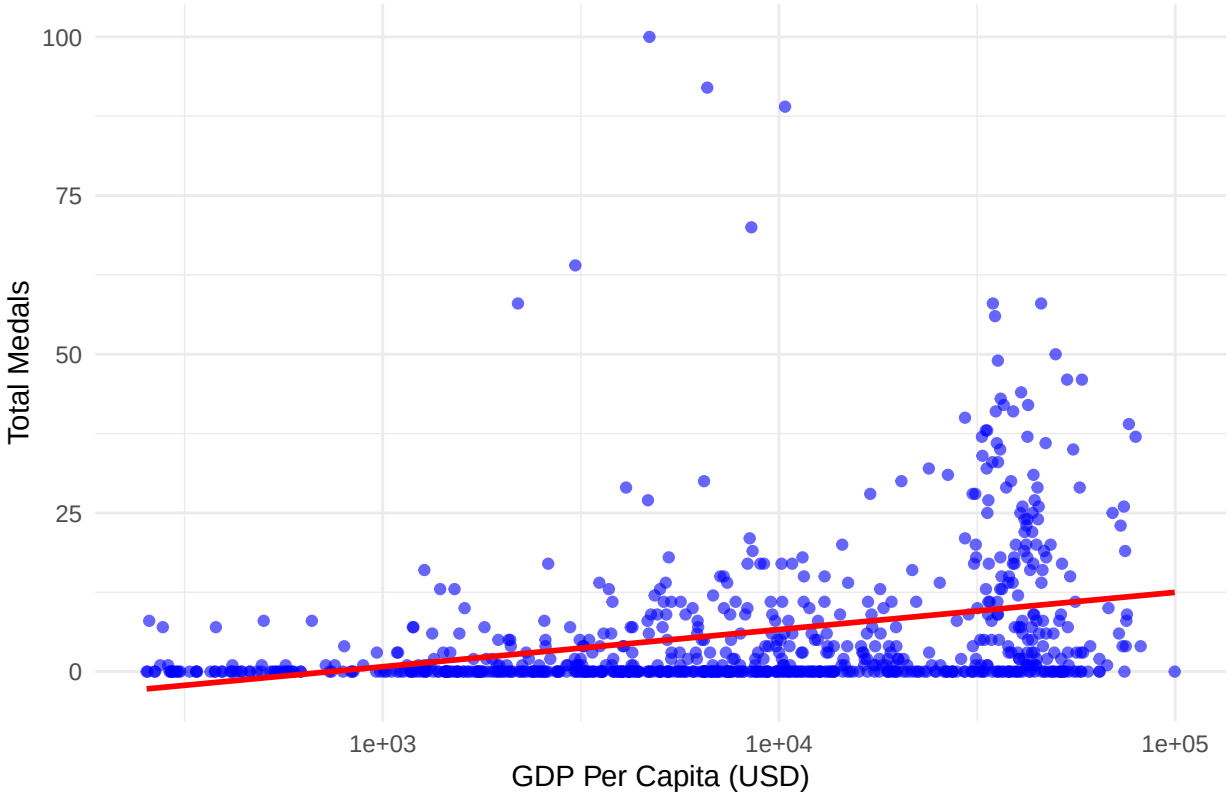

Figure 3 examines the relationship between GDP per capita and total medals won, using a log scale for GDP per capita to account for its wide distribution across countries. The positive trend in the regression line suggests that, on average, countries with higher GDP per capita tend to win more Olympic medals. However, this correlation is relatively weak, with significant variation in medal counts among nations with similar GDP per capita levels. A few wealthy nations do not perform exceptionally well, while some countries with mid-tier GDP performing extremely well. This finding suggests that while economic resources help support Olympic success, they are not the sole determinant. And we have to also think about other factors such as specialization in specific events, or population.

3.3 Comparing Multiple Economic Indicators

```
## `geom_smooth()` using formula = 'y ~ x'
```

Figure 3: Comparison of Economic Indicators vs. Medal Count

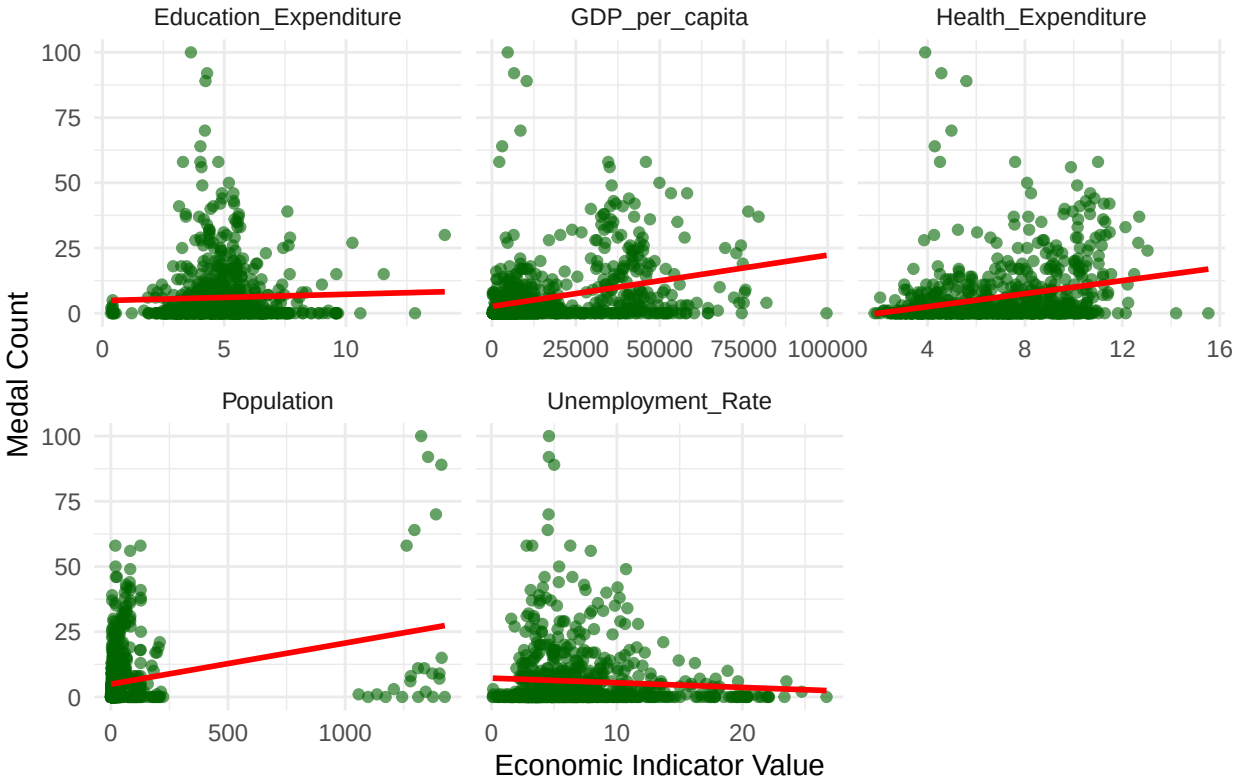

Figure 4 extends the analysis on Figure 3 by examining the relationship between total medals and multiple economic indicators. We can see the following from plots: 1. Non-logged GDP per capita continues to show a positive but weak correlation with medal counts, reinforcing the earlier findings. 2. Health expenditure is also positively correlated with medals, suggesting that countries investing in public health and well-being may indirectly support better athlete performance. This makes sense, as a healthier population could contribute to better sports participation and training outcomes. 3. Education expenditure has little correlation with almost a flat line and countries with high medal counts only spend about 5% of its GDP which is approximately the same as the median in Table 2. This indicates that a country's education budget/quality of education in a country does not correlate to Olympic success. 4. Population size has a strong positive trend between medal counts, which is expected. However, we do see a huge gap between countries with small and large populations, which is problematic. Furthemore, the countries with a large population have extremely large medal counts which could be skewing the trend. 5. We see a slight downward trend between unemployment rate and medal counts. Although this trend is very weak, we do see that countries with lower unemployment rates have higher medal counts. Furthermore, we do see that countries with high medal counts have an unemployment rate of 10% or less. As low unemployment rates are often see as an economic success, this further supports my proposed hypothesis.

3.4 Relationship Between Economic Indicators and Olympic Success

Figure 4: Correlation Matrix of Economic Indicators

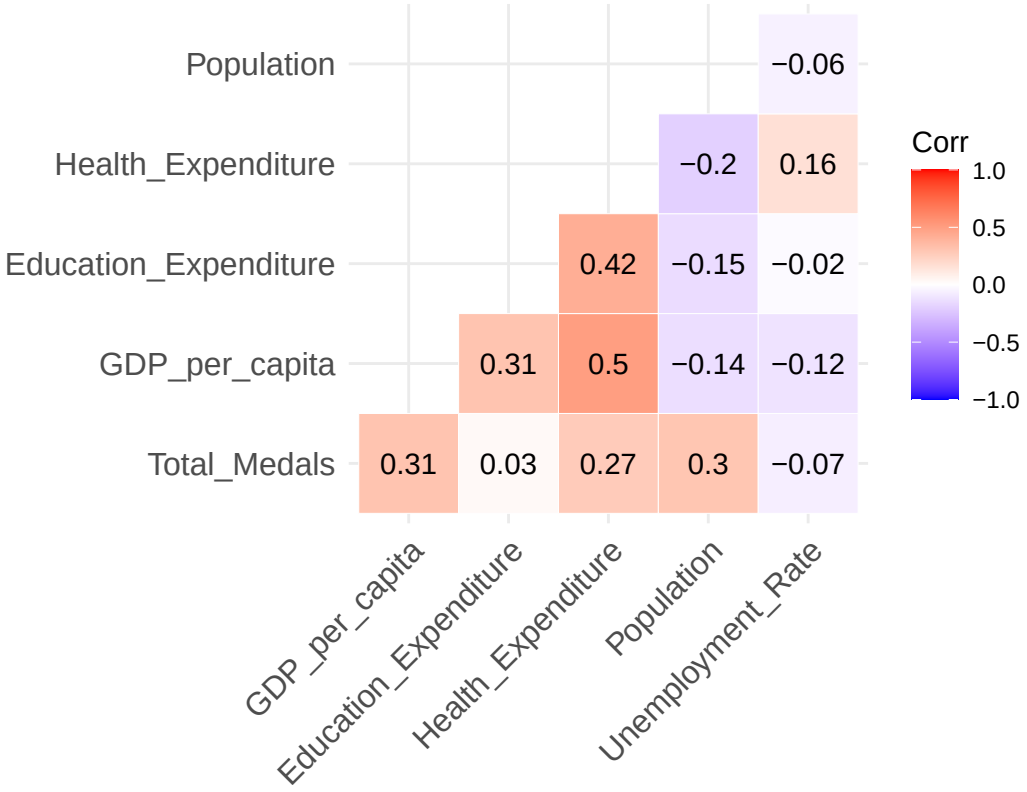

Figure 5 presents the correlation matrix of key economic indicators and Olympic medal counts. The correlation values range from -1 to 1, with positive value/red colour indicating a direct relationship, negative values/blue colour indicate an inverse relationship, and values near zero suggesting no correlation.

We see from the plot none of the variables shows an extremely strong positive correlation nor an extremely weak one. Suggesting that strong multicollinearity is not present.

Now, GDP Per Capita and Health Expenditure shows a Moderate Positive Correlation with Medal Count (r=0.31 and r = 0.27 respectively) meaning that countries with higher GDP per capita and higher health expenditure tend to win more medals, supporting the idea that economic strength contributes to Olympic success. However, this correlation is moderate, meaning that GDP and Health Expenditure does not solely determine a country's success in the Olympics.

We see that Education Expenditure (r = 0.03) has almost no correlation between the share of GDP spent on education and Olympic performance. This suggests that general education spending does not directly translate to sports success. The same can be said to unemployment rate (r = -0.07) which suggests that a country's employment status does not affect its Olympic success.

4. Summary

From the summary statistics presented, we saw that countries with better economic performance tend to have higher medal counts in the Olympics. In addition, we saw wealthier/developed countries are more present in Figure 1, further supporting the notion that countries with better economic performance have won more medals over the years. However, from figure 3 and 4 we saw that these economic factors does not fully explain how successful a country will be at the Olympics. We also saw that, within these factors some does not provide much explanation about a country's success at the Olympics.

Plans 1. Create + Train models that predicts number of Medals in the 2024 Olympics based on economic performance - Some models that I can fit: Multiple Linear Regression, Random Forest, Gradient Boosting - Potentially Include older olympic dataset to provide a larger dataset]

2. Perform Model Validation by selecting a specific olympic year's data
3. Predict 2024 Olympic Medals

- Retrieve the 2024 data from World Bank API and from the kaggle dataset.
- Generate the predicted medal counts using the data from World Bank API

4. Evaluate the Model's Performance using the kaggle dataset
5. Interpretation and Discussion of Findings + Conclusion

- Analyze which economic indicators were the strongest predictors of Olympic success.
- Investigate whether countries with similar economic profiles had similar Olympic outcomes.
- Summarize key insights on the role of economic strength in Olympic performance.