# Exploratory Data Analysis:
# 2021 Tokyo Olympics total medals

A portfolio project by Júlio Zampietro

## Key Findings

- Regarding total medals per country, there is a strong correlation with total GDP (0.831) and competing athletes (0.875). European data also shows a strong correlation between total medals and total spending on sports (0.817).

- Per capita measurements for GDP and sports spending did not correlate strongly with medal attainment, which indicates that money per athlete is more important than money per citizen. This provides some additional advantage to larger countries.

- Countries that had an over-average (see documentation) quantity of athletes competing in collective sports tend to fare better. This is counterintuitive at first glance, as collective sports offer a "less dense" medal-opportunity-to-athlete ratio. However, the correlation with total spending provides a basis for arguing that, while individual sports are less costly for each competing country, countries that can afford to also invest in collective sports excel regarding total medals.

- In conclusion, we recommend that countries increase total spending on both categories of sports in order to increase total athletes and, consequentially, total medals.

## Building a dataset & correlation matrixes

Data was collected from various sources to build an initial dataset with country-level data on total medals, efficiency (see documentation), total athletes, total GDP, per capita GDP, total education spending, per capita education spending, and GINI coefficient. A correlation matrix (absolute values) was created:

| | Total_medals | Efficiency | Total_Athletes | Total_GDP | GDP_per_capita | Total_edu_spending | Per_capita_edu | Gini_coefficient |
|---|---|---|---|---|---|---|---|---|
| **Total_medals** | 1 | 0,3873 | 0,8751 | 0,8314 | 0,2387 | 0,7705 | 0,1622 | 0,1873 |
| **Efficiency** | 0,3873 | 1 | 0,0857 | 0,2616 | 0,0271 | 0,2565 | 0,1888 | 0,1635 |
| **Total_Athletes** | 0,8751 | 0,0857 | 1 | 0,6678 | 0,2976 | 0,5467 | 0,2375 | 0,1078 |
| **Total_GDP** | 0,8314 | 0,2616 | 0,6678 | 1 | 0,1568 | 0,9914 | 0,019 | 0,0212 |
| **GDP_per_capita** | 0,2387 | 0,0271 | 0,2976 | 0,1568 | 1 | 0,0186 | 0,8681 | 0,3756 |
| **Total_edu_spending** | 0,7705 | 0,2565 | 0,5467 | 0,9914 | 0,0186 | 1 | 0,0227 | 0,051 |
| **Per_capita_edu** | 0,1622 | 0,1888 | 0,2375 | 0,019 | 0,8681 | 0,0227 | 1 | 0,2636 |
| **Gini_coefficient** | 0,1873 | 0,1635 | 0,1078 | 0,0212 | 0,3756 | 0,051 | 0,2636 | 1 |

A second, more focused, dataset was created for European data, seeing as it was the only region from which it was relatively easy to find data on sports investment (total and per capita):

| | Total_medals | Efficiency | Total_Athletes | Total_GDP | GDP_per_capita | Total_edu_spending | Per_capita_edu | Gini_coefficient | Total_sport_spending | Eurosport_per_capita |
|---|---|---|---|---|---|---|---|---|---|---|
| **Total_medals** | 1 | 0,4355 | 0,9204 | 0,8263 | 0,0993 | 0,809 | 0,0138 | 0,016 | 0,8171 | 0,1962 |
| **Efficiency** | 0,4355 | 1 | 0,147 | 0,104 | 0,1346 | 0,1069 | 0,0312 | 0,1668 | 0,1137 | 0,0865 |
| **Total_Athletes** | 0,9204 | 0,147 | 1 | 0,9051 | 0,1331 | 0,8908 | 0,0111 | 0,0579 | 0,8997 | 0,1442 |
| **Total_GDP** | 0,8263 | 0,104 | 0,9051 | 1 | 0,1863 | 0,9866 | 0,0829 | 0,1056 | 0,9246 | 0,1482 |
| **GDP_per_capita** | 0,0993 | 0,1346 | 0,1331 | 0,1863 | 1 | 0,2699 | 0,9028 | 0,2475 | 0,1203 | 0,5005 |
| **Total_edu_spending** | 0,809 | 0,1069 | 0,8908 | 0,9866 | 0,2699 | 1 | 0,2027 | 0,0723 | 0,9241 | 0,2441 |
| **Per_capita_edu** | 0,0138 | 0,0312 | 0,0111 | 0,0829 | 0,9028 | 0,2027 | 1 | 0,1962 | 0,0682 | 0,6927 |
| **Gini_coefficient** | 0,016 | 0,1668 | 0,0579 | 0,1056 | 0,2475 | 0,0723 | 0,1962 | 1 | 0,0532 | 0,3007 |
| **Total_sport_spending** | 0,8171 | 0,1137 | 0,8997 | 0,9246 | 0,1203 | 0,9241 | 0,0682 | 0,0532 | 1 | 0,0187 |
| **Eurosport_per_capita** | 0,1962 | 0,0865 | 0,1442 | 0,1482 | 0,5005 | 0,2441 | 0,6927 | 0,3007 | 0,0187 | 1 |

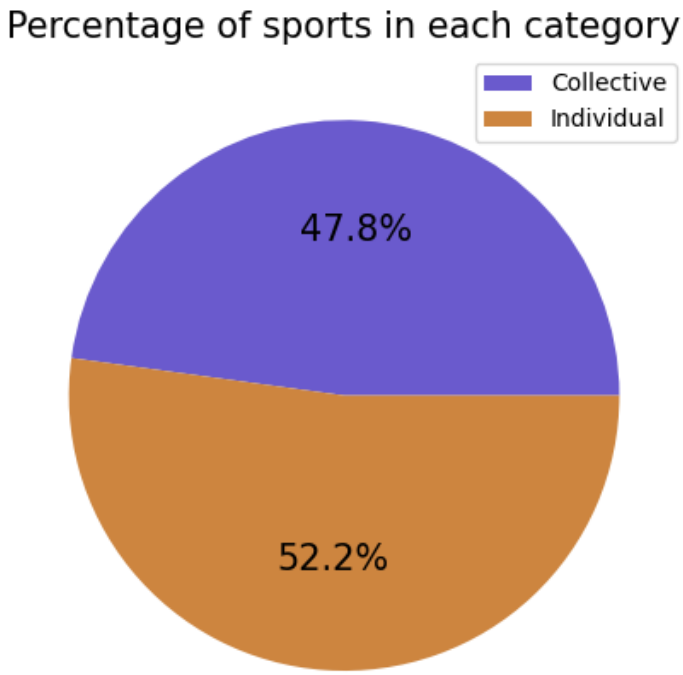
- Total_medals strongly correlates to Total_Athletes, Total_GDP, Total_edu_sending, and Total_sport_spending, and does not correlate well with the per capita variants of these metrics.

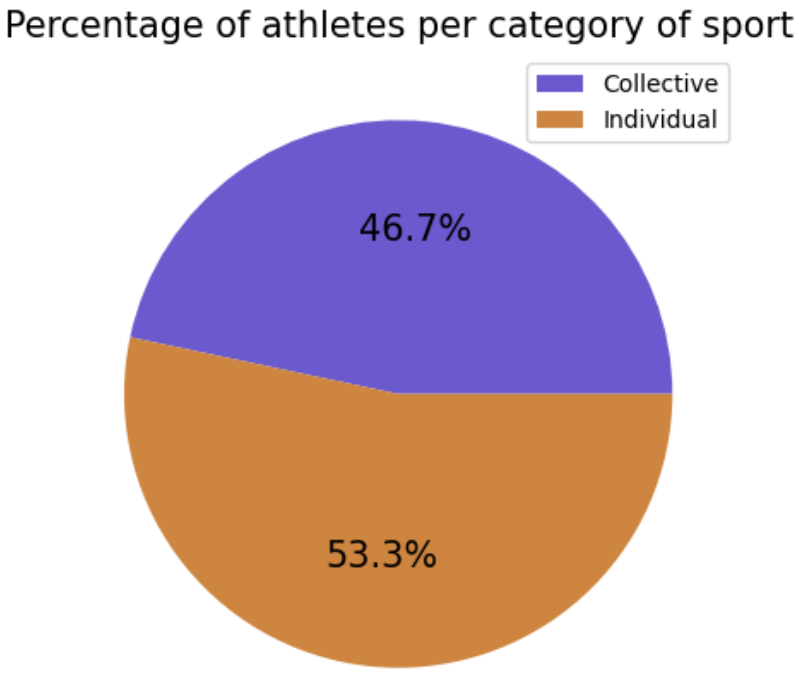
- As Total_edu_spending and Total_GDP correlate almost perfectly (0.991 for general data), we assume that the former is a function of the latter, so that only Total_GDP is relevant.

- Total_sport_spending also strongly correlates with Total_Athletes (0.899).

**Individual vs. Collective sports**

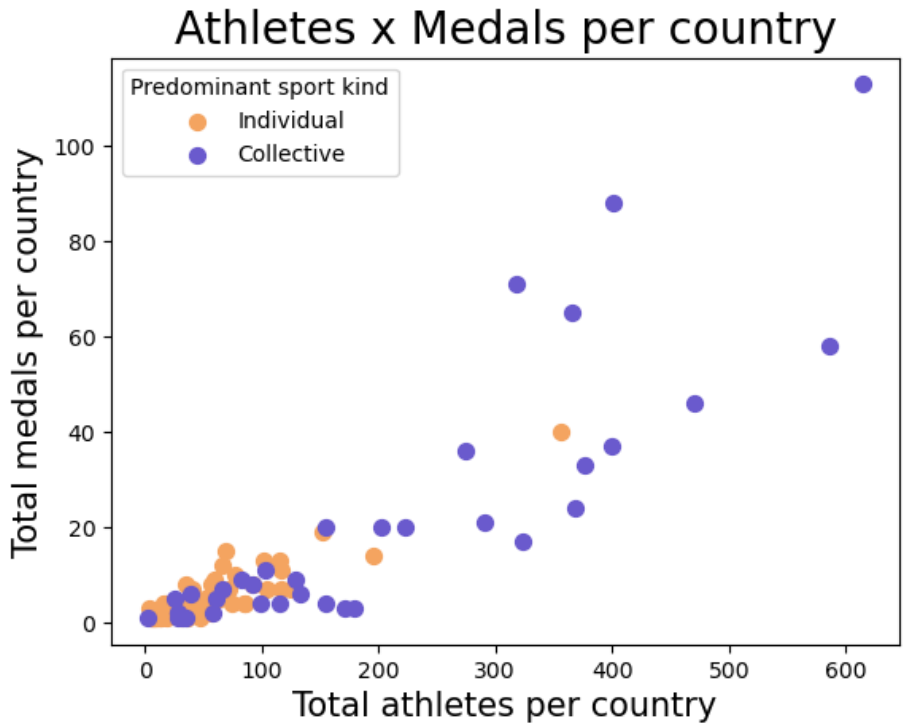
- Used SQL to obtain the total number of athletes competing in collective (5174, 46.67%) and individual (5911, 53.32%) sports.
- Used SQL to create a sports classification table, dividing sports into collective or individual. Both sports and athletes distribute somewhat equally between categories:



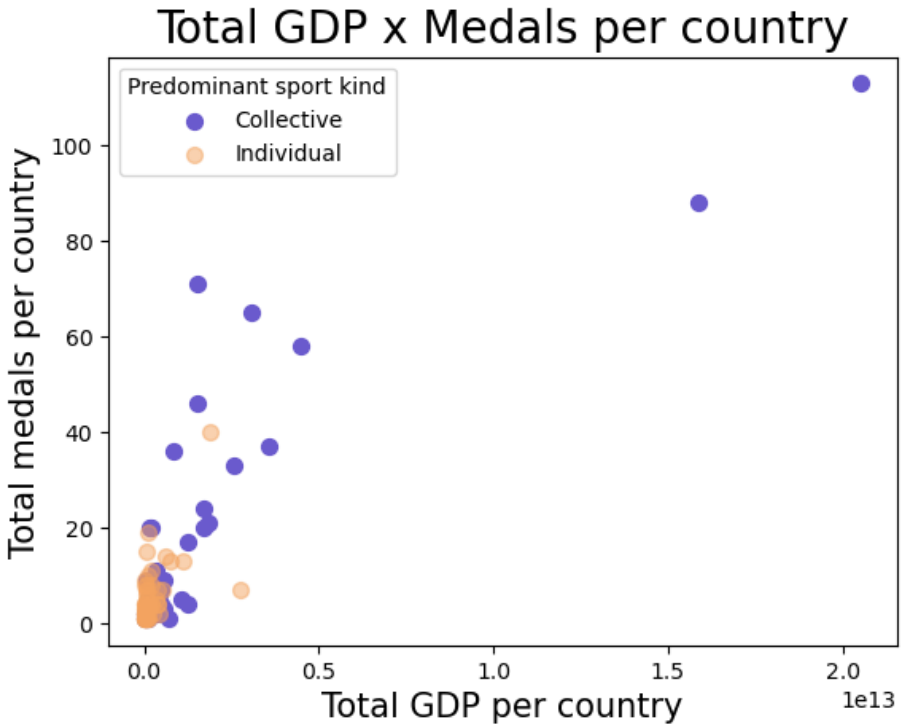



- A clear pattern emerges regarding individual and collective sports when plotting the three most relevant variables:

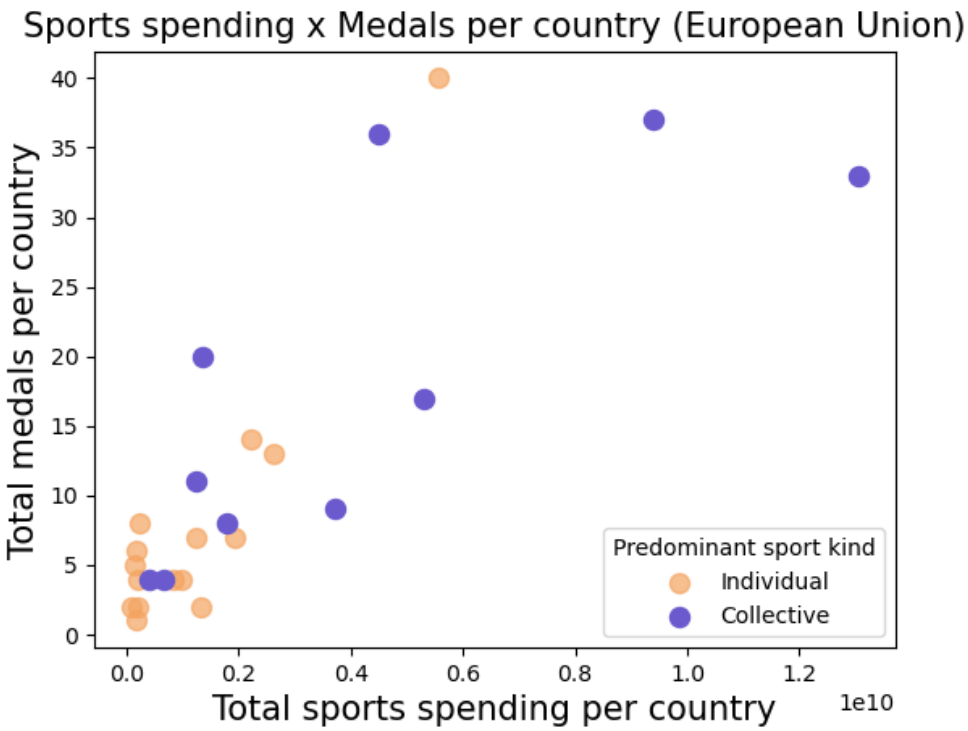
**Graph 1**




**Graph 2**



**Graph 3**




- While countries with primarily individual athletes stick to the bottom of the total medal rankings, those with primarily collective athletes tend to reach better positions.

- The pattern is not explained by athlete distribution between both categories, as it is very similar to the distribution of sports. It is also not explained by athlete efficiency (see documentation) as noted in the low correlation index.

- One plausible explanation is that while individual sports require the least amount of total investment to yield medals, countries that have the economic resources to invest in both individual and collective categories tend to excel in aggregate.

- We conclude that one way to increase total medal yields is to increase sports investments across the board (both categories), which will lead to an increase in athlete participation (0.899 correlation) and, consequentially, an increase in total medals.