Data Wrangling
Project report 2023

What is the correlation between a country's various indexes and the amount of medals awarded during the Olympic Games?

Project group members

Karolina Hajkova, 2728951, kha520
Matt Hrkal, 2737442, mhl227
Antal Deurloo, 2706710, ado660
Malgorzata Zdych, 2730740, mzh800

Research question

What is the correlation between a country's various indexes and the number of medals awarded during the Olympic Games? How do different statistics of countries correlate to how successful they perform at the Olympic Games?

The factors that have been tested for correlation with success on Olympic Games are as follows;

- Corruption Perceptions Index (CPI)
- Alcohol consumption
- Happiness score
- Population density
- Median Age
- Urban Population

Data sources

Kaggle online database was used to fetch the data on the medal winners, population of countries and all the factors that were tested whether they correlate with the success of countries at the latest summer Olympic Games.

Tokyo 2020 Olympics Medals (Kaggle):
https://www.kaggle.com/datasets/berkayalan/2021-olympics-medals-in-tokyo [23.01.2023]

Population by country 2020 (Kaggle):
https://www.kaggle.com/datasets/d893bc6eb4370c2fd7b87bcf41972963b607202a1683f576700c52e6ecd4ab2a [25.01.2023]

Global Corruption Perceptions Index for 10 years (Kaggle):
https://www.kaggle.com/datasets/jeegarmaru/corruption-perceptions-index-for-10-years?select=CPI_2019_final_dataset.csv [23.01.2023]

Alcohol consumption around the world (Kaggle):
https://www.kaggle.com/datasets/codebreaker619/alcohol-comsumption-around-the-world/download?datasetVersionNumber=1 [01.02.2023]

Population by Country (Kaggle):
https://www.kaggle.com/datasets/tanuprabhu/population-by-country-2020/download?datasetVersionNumber=4 [01.02.2023]

Data wrangling methods

The three datasets that were initially found included the ranking of 2020 Tokyo Olympics Medals per country, Global Corruptions Perceptions Index (CPI) for the last 10 years and Population by country in 2020. The data was consecutively presented in a medals.csv file (92 countries), cpi.csv (179 countries) and population.csv (234 countries). The pandas library was applied to extract and read the information from Comma Separated Values (CSV) data format. Firstly, the problem across the files was recognised as the different naming conventions between the countries' names, for instance China in cpi.csv and People's Republic of China in medals.csv. Therefore, all the names needed to be standardized before the merging process. The function normalize_function takes as an argument dataframe and the name of the column with country name was created, it implements “pycountry”, “pycountry_convert” and “fuzz” with “process” from “fuzzywuzzy” library that is specific for handling non-standard data. The function was supposed to find the matches between the naming conventions of a given data frame and a standardized set of country names as defined by “pycountry” library. The match value was set to 74, after evaluating the string similarity scores of the countries that were caught during the first exception. Moreover, in the case of medals.csv there was no column that represented the ISO 3166-1 alpha-3 country code, thus an additional translation from the original country name to the three letter abbreviations was performed, for example from China to CHN. Moreover, few instances of countries needed to be mapped to their alpha-3 country codes manually - Great Britain, Russia and Taiwan. These were special cases that had linguistically very different variations. The fuzzywuzzy library was unable to find a connection between strings "Great Britain" and "United Kingdom" due the diverse variations of the name. On top of that, Russian and Taiwanese athletes could not represent their countries at the Olympic Games and hence took part under ROC and NOC. The team has decided to map these as equivalents despite the political discrepancy. Additionally, Kosovo does not have their ISO 3 code at all, which led to its exclusion from the analysis as either way data for the different variables was not available for Kosovo.

Eventually, the data frame df_medals that was populated using medals.csv was transformed and displayed with the country codes that replaced actual names of the countries. Furthermore, the new methodology was implemented and once the non-standard data was handled in df_medals, the medal scoring system was developed. It was decided that Golden medals correspond to 3 points, Silver ones to 2 points and Brown to 1 point. According to this scheme, the values of medals were recalculated in the data frame and included in the analysis. Subsequently, a new df_medals data frame was merged with the df_population dataframe based on the CCA3, also known as ISO3 - three letter code. The dataset was adjusted to be more representative in the visualizations, therefore medal value per million inhabitants was determined, and saved as data frame “medals_per_million”. This final adjusted value of the value of the medals won per million inhabitants of a country was intended to objectively evaluate how successful each country performed at the last summer Olympic Games.

Finally, the resulting data frame of medals and population was one by one merged with the CPI, Alcohol consumption, Happiness score, Population density, Median Age, and Urban Population. The merging type chosen was LEFT JOIN with the intention to preserve as many countries as possible for each factor that will be evaluated. The LEFT JOIN fills all missing value from the records of the right data frame that are included in the result with NaN. This practice preserves more rows in the dataframe we are working with.

Value cleaning

Compared to the naming normalization, the other data cleaning was quite straightforward. A few column names needed to be changed for readability, some columns' data types had to be changed from objects to numerical values. The pandas functions 'rename()' and 'astype()' served for this purpose. In the case of the urbanization grades the values needed to be cleaned first. These values included a space and a '%' character, which we removed using the following regex function:

```
df.replace(' %', '', regex=True, inplace=True)
```

The NaN values also needed to be removed for the 'astype()' operation. Luckily, most countries where NaN values were present did not score any medals, thus were not usable anyway. Furthermore, the NaN values in medal-scoring country rows were often in the same countries, where it is tough to gather data. Thus, the rows containing NaN values could be removed with only a slight loss to the dataframe.

We dropped two countries from the dataset after classifying them as outliers because of their extreme values. The Bahamas, which scored a whopping 14 medals per million inhabitants, and Bahrain, which, at 2239(P/KM²), greatly exceeded the population density of all other included countries.

The data could now easily be plotted using matplotlib. The trend lines were added with the following code:

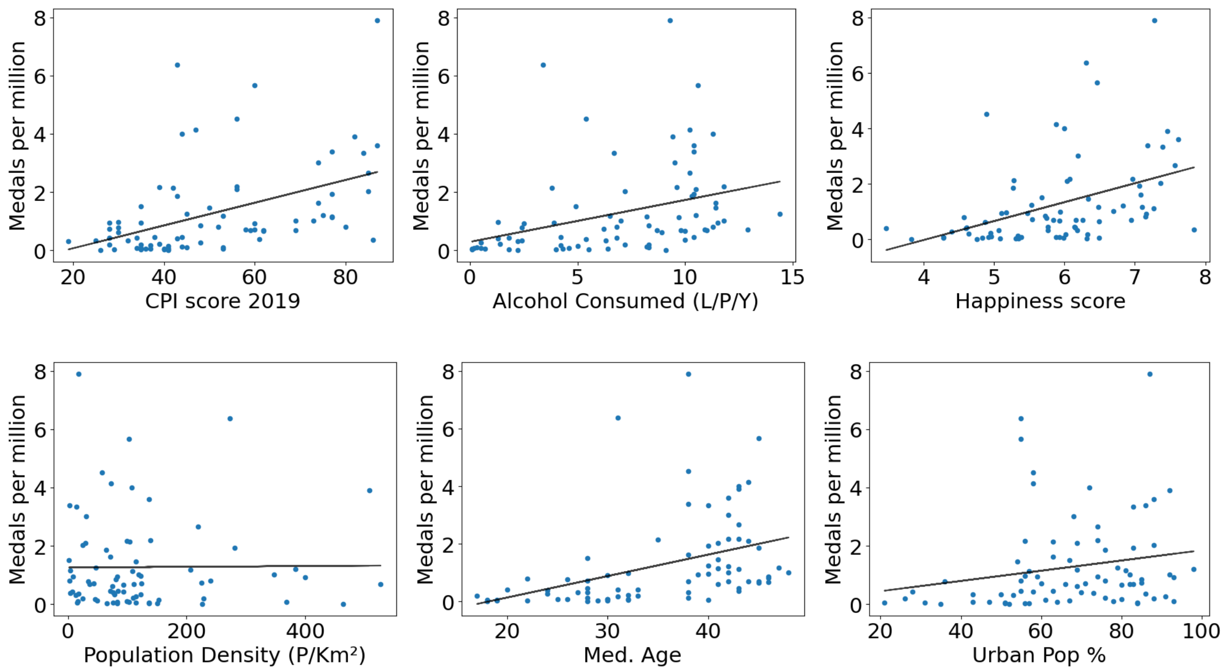
```
z = np.polyfit(df[x_data], df[y_data], 1)

p = np.poly1d(z)

plt.plot(df[x_data],p(df[x_data]), 'k', alpha=0.75)
```

Analysis

Graphs showing the trend lines between the tested factors and the medal value per million inhabitants:

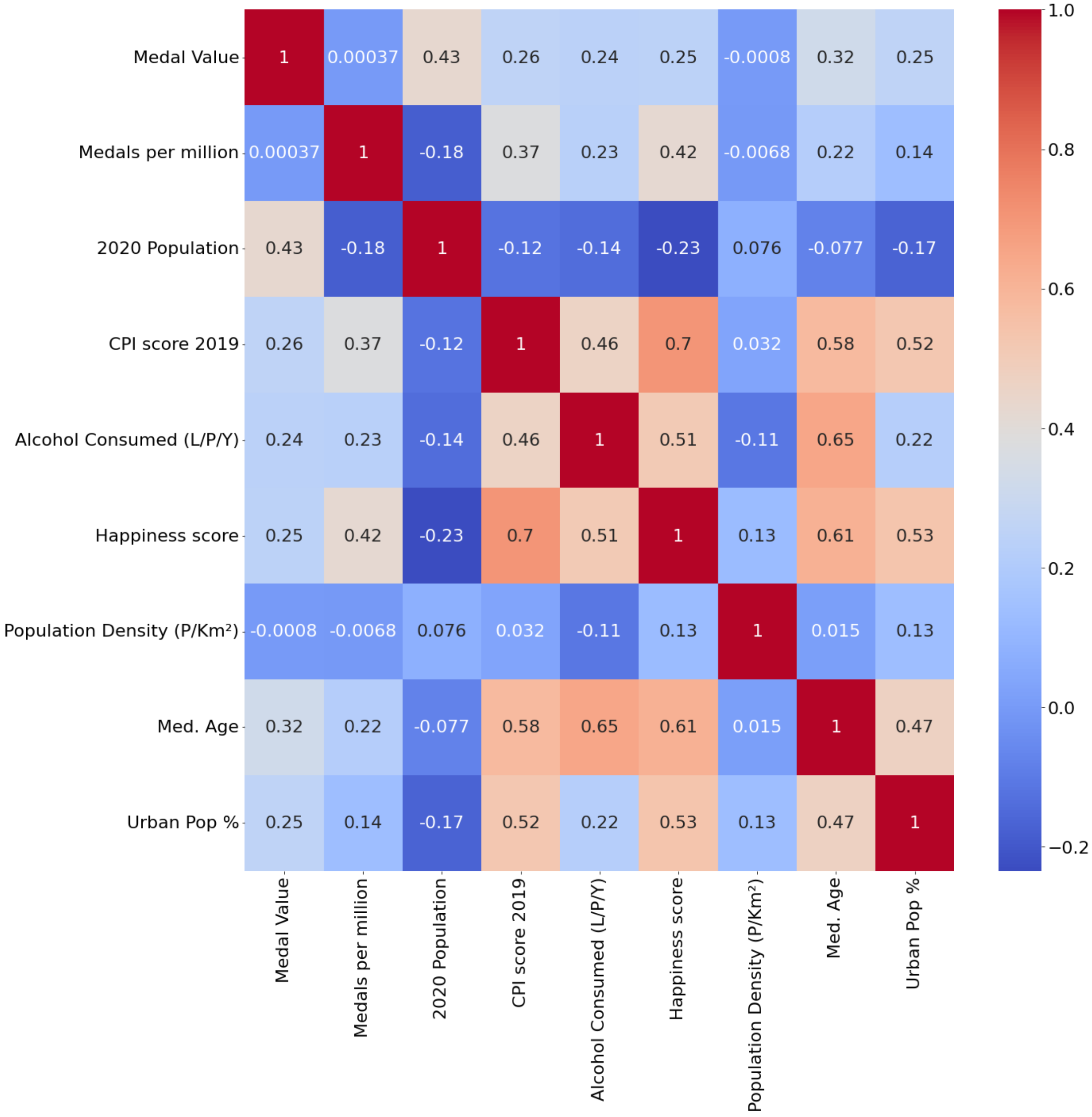


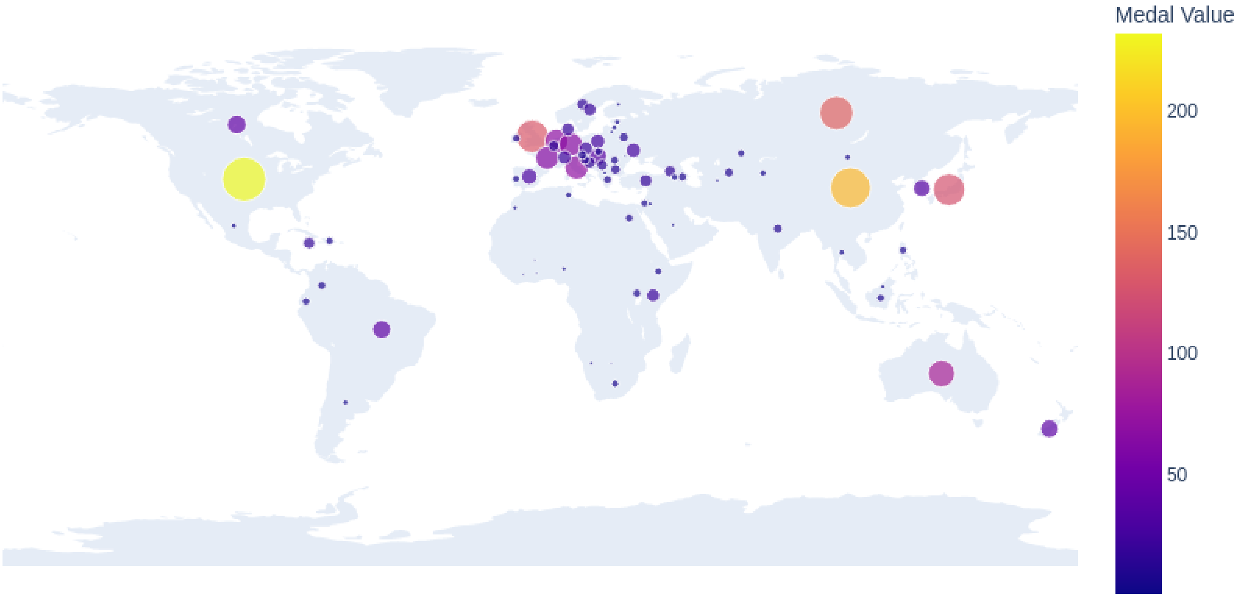
Correlation coefficients between all considered factors:

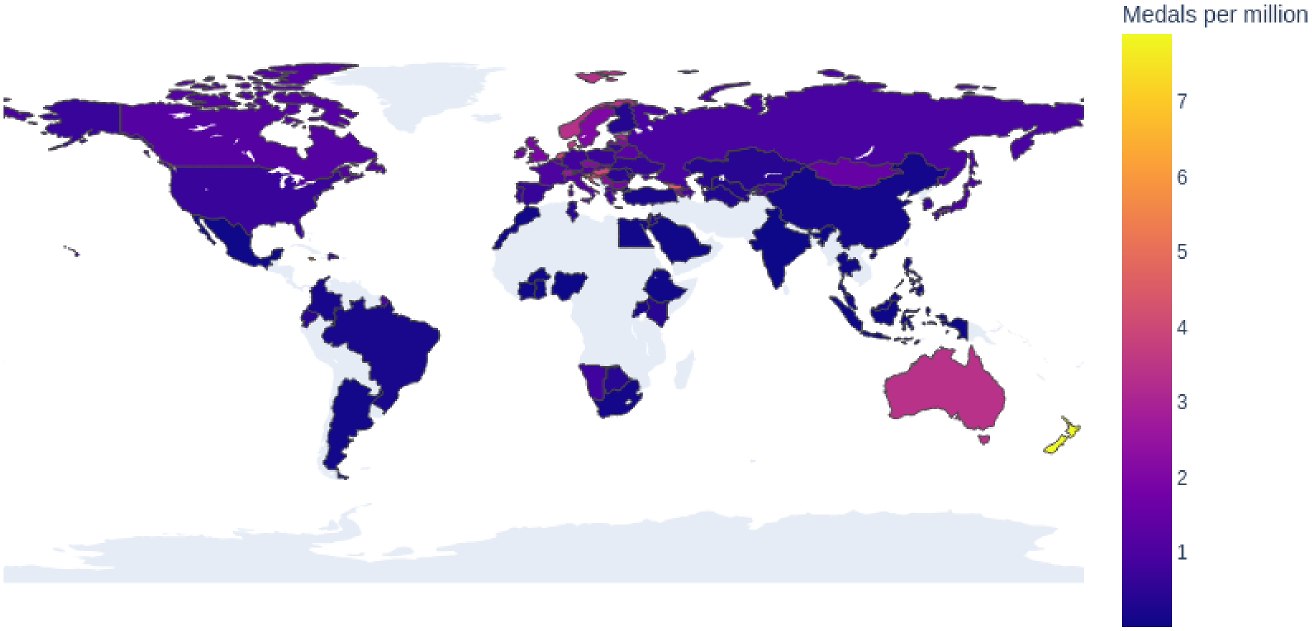
Population & Medal value = 0.43: This suggests a moderate positive correlation between population and medal value, meaning that as the population of a country increases, the medal value (the total number of medals won) also tends to increase. This could be because larger countries have a larger pool of athletes to select from and more resources to allocate to training and competing in the Olympics.

Medals per million & CPI Score = 0.37: This suggests a moderate positive correlation between medals per million and the Corruption Perceptions Index (CPI) score. This means that countries with higher medals per million (i.e., countries that are more successful in the Olympics on a per capita basis) tend to have higher CPI scores (i.e., they tend to be perceived as less corrupt). This could be because countries that invest more in their athletes and sporting infrastructure may also invest more in other areas, such as governance and rule of law, that are related to lower corruption.

Medals per million & Happiness score = 0.42: This suggests a moderate positive correlation between medals per million and happiness score, meaning that countries that are more successful in the Olympics on a per capita basis tend to have higher levels of reported happiness. This could be because sporting success is a source of national pride and can contribute to a sense of collective achievement and well-being.


Medal Value
1
0.00037
0.43
0.26
0.24
0.25
-0.0008
0.32
0.25
Medals per million
0.00037
1
-0.18
0.37
0.23
0.42
-0.0068
0.22
0.14
2020 Population
0.43
-0.18
1
-0.12
-0.14
-0.23
0.076
-0.077
-0.17
CPI score 2019
0.26
0.37
-0.12
1
0.46
0.7
0.032
0.58
0.52
Alcohol Consumed (L/P/Y)
0.24
0.23
-0.14
0.46
1
0.51
-0.11
0.65
0.22
Happiness score
0.25
0.42
-0.23
0.7
0.51
1
0.13
0.61
0.53
Population Density (P/Km²)
-0.0008
-0.0068
0.076
0.032
-0.11
0.13
1
0.015
0.13
Med. Age
0.32
0.22
-0.077
0.58
0.65
0.61
0.015
1
0.47
Urban Pop %
0.25
0.14
-0.17
0.52
0.22
0.53
0.13
0.47
1
Medal Value
Medals per million
2020 Population
CPI score 2019
Alcohol Consumed (L/P/Y)
Happiness score
Population Density (P/Km²)
Med. Age
Urban Pop %
1.0
0.8
0.6
0.4
0.2
0.0
−0.2





An interesting comparison is to use medal values. We used both bubble and choropleth maps to compare the number of medals discounted by population size to the artificial metric of Medal Value. On the surface, it can appear that countries with larger populations have a higher Medal Value, and thus, more medals in total. However, when we look closer at the data, we can see that, although smaller countries such as New Zealand have a lower Medal Value, meaning they have fewer medals with combined scores on the lower end of the spectrum, they are actually on top when looking at the number of medals per million inhabitants. This is a significant finding, as it demonstrates that size doesn't necessarily equate to success.

It is important to consider the context of the analysis, as we have no information on whether the medals are coming from a few athletes or distributed across multiple. This is one drawback of our analysis and something we are not able to tell with the current combined dataset. What we can tell, however, is that smaller countries have the potential to succeed and make an impact, even if their population size is comparatively small. This is a significant insight that can help inform policy decisions and give smaller nations a chance to create a better environment for athletes.

Conclusion

It's worth noting that correlation does not necessarily indicate causality, meaning that it's plausible that the relationships observed in this analysis may not be causal in nature. Therefore, further investigation is needed to confirm the causal nature of the relationships, such as through the use of causal inference methods. Additionally, it's important to consider other potential confounding factors that may influence these relationships, such as income, education levels, and cultural attitudes toward sports. These factors may explain a portion of the correlation observed and should be taken into account when interpreting the results. Despite these limitations, this analysis provides valuable insights into possible relationships between various variables and highlights the importance of considering multiple factors when exploring correlations. Furthermore, this analysis can serve as the basis for further research, allowing us to better understand the relationship between different variables and to develop more accurate predictive models.