Visual Data Science

Wrangler and Profile

Clara Pichler[1]

1 Wrangler

In this section, I will describe the preprocessing steps required for merging the datasets. These steps include formatting, handling missing values, and dropping unnecessary columns. While I cannot detail every step due to the extent of work involved, I will highlight the most critical aspects.

For the `df_olympics` dataframe, unnecessary columns such as `athlete_url`, `country_code`, and `country_3_letter_code` were dropped. Initially, I planned to merge using the ISOCode, assuming it to be `country_3_letter_code`, but I found discrepancies (e.g., Zimbabwe should be ZWE but is listed as ZIM). The `slug_game` attribute was modified to `year`, leaving out the host country as it was irrelevant to this project. The `year` column was converted to `int`, while other types remained unchanged. I considered converting `medal_type` (gold, silver, bronze) to numeric values (1, 2, 3) for easier analysis in statistical or machine learning models. Additionally, the `athlete_full_name` column was dropped, which resulted in duplicate rows for team events (`participant_type`: `GameTeam`). To resolve this, I applied `drop_duplicates()`. For simplicity, only Olympic games from 1960 to 2022 were included.

For `df_elevation`, I dropped columns like `Highest point` and `Lowest point`, renaming the remaining ones for clarity. The country column was crucial, as it was used for merging. Elevation-related columns (`max_elevation`, `min_elevation`, `elevation_span`) were reformatted to integers, replacing `sea level` with 0. Since the dataset was scraped from a website, such formatting adjustments were necessary. Similar steps were taken for `df_climate`, which provides climate zone information. Unnecessary columns were dropped, and value types were cleaned. The GDP dataset required extensive restructuring, as it presented each year as a column. The final `df_gdp` dataset consisted of `country_name`, `isocode`, `year`, and `gdp` columns.

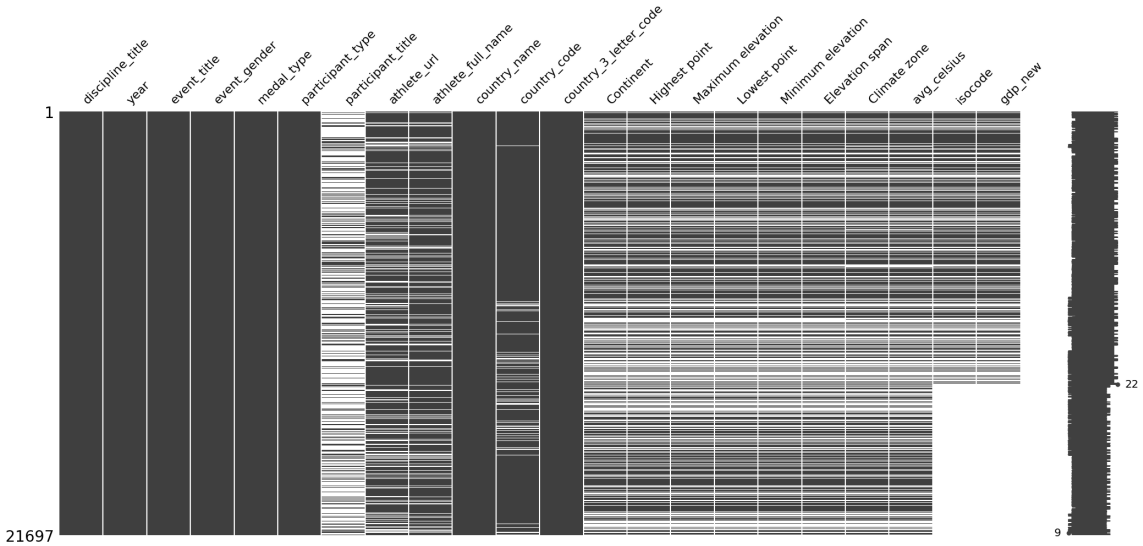
Handling missing values was divided into two stages: before and after merging. Before merging, I used methods like linear interpolation for GDP. After merging, further imputations were necessary, sometimes utilizing data from other datasets. Some countries had different naming conventions across datasets, requiring renaming. Additionally, countries no longer existing, such as Czechoslovakia, Serbia and Montenegro, Soviet Union, and Yugoslavia, posed challenges. For these, I researched constituent countries to estimate values: `elevation` was filled with the minimum (maximum) of their respective values, `climate` averaged over Celsius, and climate zones determined by the median. Specific Olympic teams, like the Unified Team and Independent Olympic Athletes, were handled based on their participating years and the athletes' associated countries (e.g., Serbia in 1992 and Kuwait in 2016 for Independent Olympic Athletes).

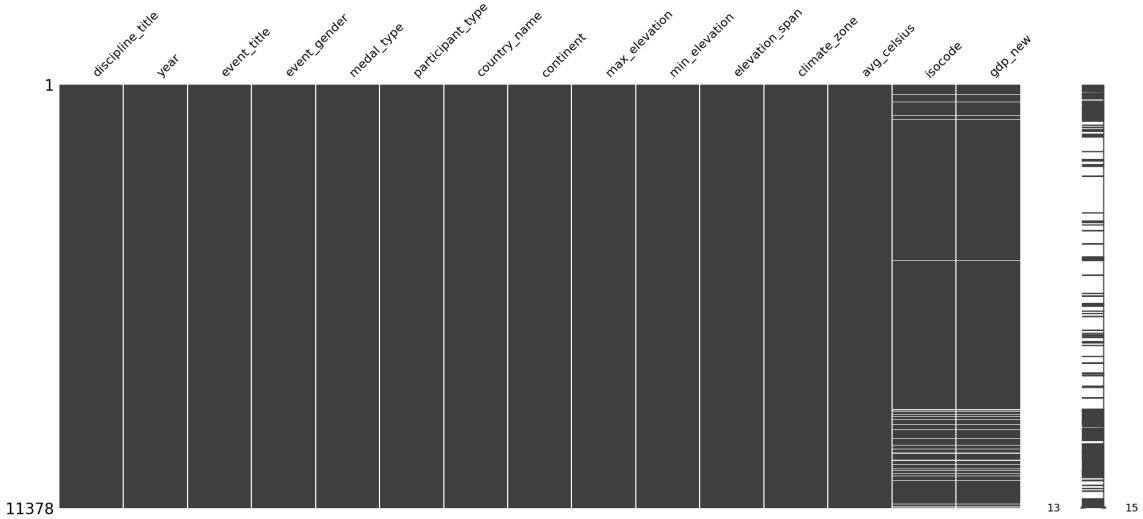
For merging, I performed a left join of `df_olympics` with `df_elevation` and `df_climate` using the key `country_name`, and with `df_gdp` using `country_name` and `year` as keys.

To monitor missing values throughout, I used the `msno` matrix. Figure 1 shows the merged dataset before and after preprocessing (renamed keys were necessary for merging). Despite these efforts, some missing values remain. For example, countries participating in the Olympics but missing GDP data (e.g., due to absence from the World Bank dataset) still pose challenges. Suggestions

[1] 11917694 - e11917694@student.tuwien.ac.at

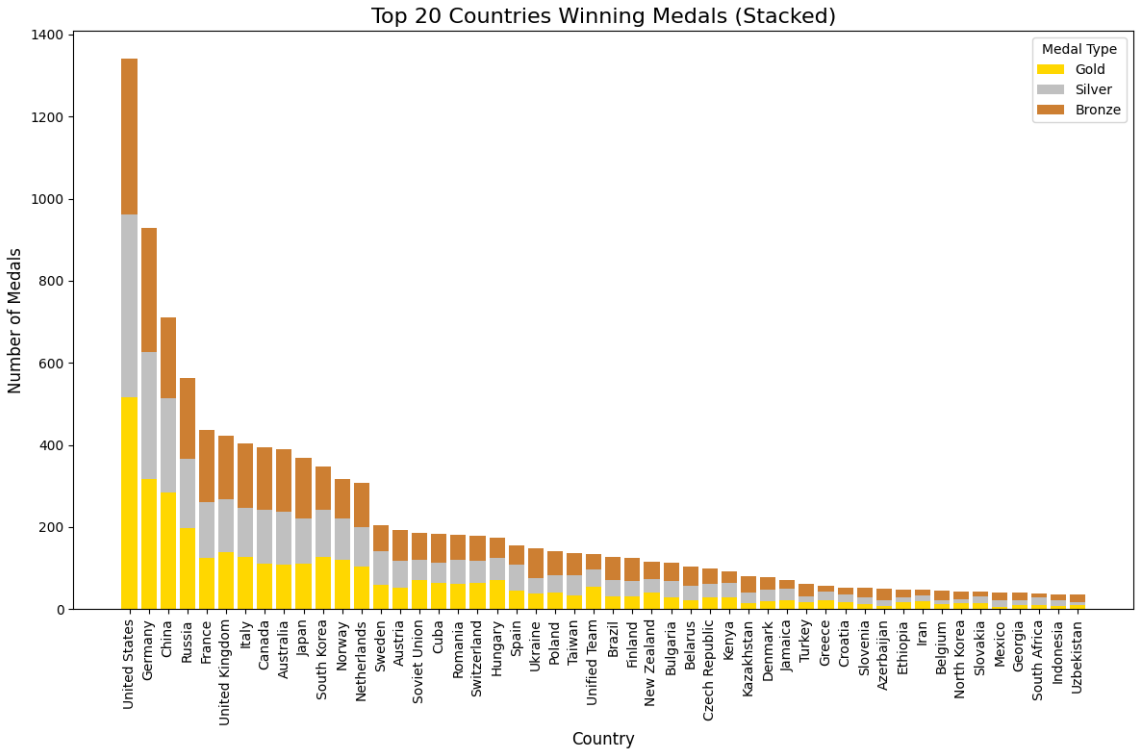
for handling this are welcome. While manually adding GDP values from sources like Wikipedia is beyond this task's scope, I could consider excluding these countries or omitting GDP as a feature, though both options would result in significant data loss. Feedback is appreciated.





2 Profile

2.1 Top 50 Countries Winning Medals

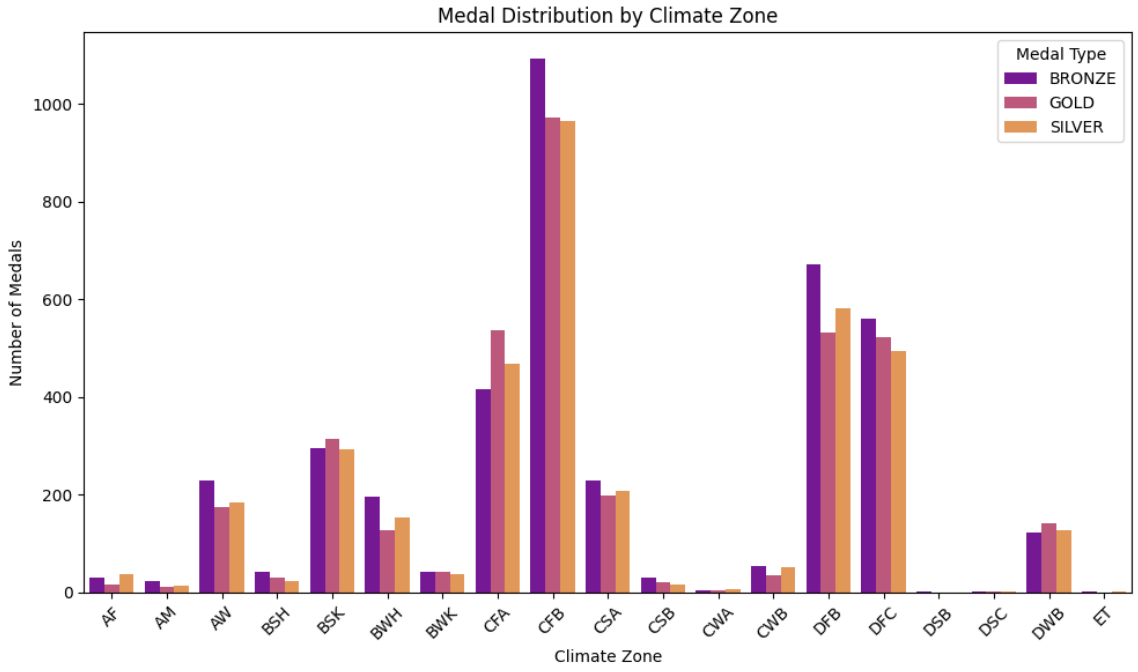


The stacked bar chart provides a clear visualization of the total number of Olympic medals (gold, silver, and bronze) won by leading nations. It highlights the dominance of a few countries. The United States is the most dominant nation, with a significantly larger total medal count compared to all other countries. The bar for the United States stands out dramatically, reaching nearly 1,350 medals. This can be attributed to its strong athletic programs, extensive infrastructure, and investment in sports development. Germany, China, and Russia follow with total medal counts significantly lower but still a lot compared to others. Germany's performance is particularly notable, especially when compared to China, which has seen rapid population growth in recent decades. These countries' successes may reflect factors such as sports culture, state support, and population size.

A striking insight is the rapid decline in medal totals after the top 10 countries. The countries from South Korea onward show much smaller bar heights, indicating that a relatively small group of nations consistently dominate international competitions. Additionally, countries like Netherlands, Sweden, and Norway show proportional balance between their gold, silver, and bronze medals.

The visual emphasizes the disparity in sports performance globally, with a small number of nations accounting for the vast majority of medals. This may be tied to factors like economic investment in sports, infrastructure, and national priorities. While countries lower on the chart still achieve significant success, they are overshadowed by powerhouse nations like the United States and Germany.

2.2 Medal Distribution by Climate Zone



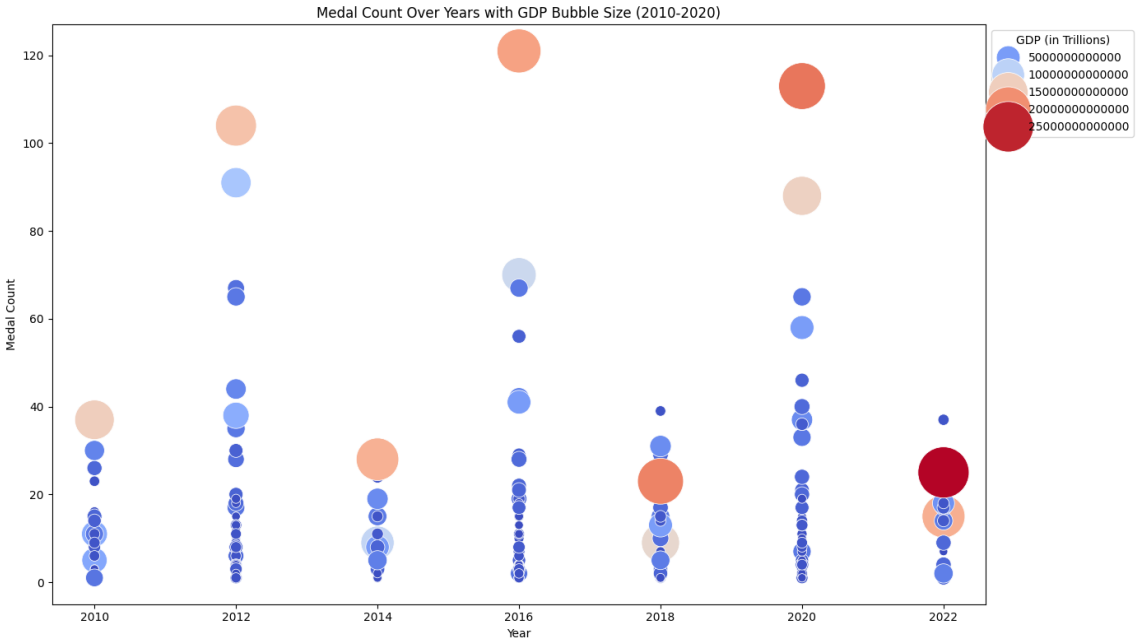
The bar chart 2.3 provides insights into how Olympic medals are distributed across various climate zones. This graph highlights notable disparities between climate zones, suggesting potential links between geography, climate, and athletic success.

The CFB climate zone (warm temperate with no dry season and warm summer) is the most dominant by a significant margin, with over **1,000 medals**. Interestingly, the distribution of gold, silver, and bronze medals in this zone is nearly balanced, indicating consistent performance across all medal types. The dominance of this climate zone may be attributed to countries with significant sporting histories, such as parts of Europe (another bar plot which is not here included shows that Europe has by far the most medals) and regions like the United Kingdom and Australia, which fall into this category.

Other climate zones with notable medal counts include CFA (humid subtropical), DFB (cold climate with no dry season and warm summer) and DFC (cold climate with no dry season and cold summer). CFA shows significant gold medal achievements compared to bronze and silver, suggesting strength in top performance. Similarly, DFB features a high overall medal count, with slightly more bronze medals, indicating strong competitive representation from countries in colder climates, such as Russia, Canada, and parts of Northern Europe. Medal counts in these colder zones (DFB/C) may reflect the prevalence of winter sports, where countries with cold climates typically excel.

In contrast, climate zones such as AF (tropical rainforest), AM (tropical monsoon), and BWK (cold desert) show much smaller medal counts. These regions have minimal representation, suggesting that geographic, infrastructural, or socioeconomic factors might limit athletic participation and success on the global stage.

2.3 Medal Count Over Years with GDP (2010-2020)



This bubble graph visualizes the relationship between a country's medal count and its GDP from 2010 to 2022. A clear trend emerges: countries with higher GDPs tend to win more medals. This suggests that economic resources play a significant role in supporting high-performance sports programs and infrastructure. However, it's important to note that this correlation is not absolute. Factors like government policies, cultural emphasis on sports, and individual athlete talent also contribute to a country's sporting success.

While the graph reveals a general trend, it's also interesting to observe individual country performances. China consistently stands out with both a high medal count and a large GDP, reflecting its significant investment in sports development. The United States, with its established sporting culture and resources, also maintains a strong presence. In contrast, Russia's medal count has been impacted by recent doping scandals and sanctions.