CIS 9650 Project Group 3 Final Project Write Up December 18, 2021

Team & Tasks:

Team Members	Role
Ingrid Becker	Python Programming
Laye Oumar Koulibaly	Python Programming & Data Cleanup
Luke Liao	Database Programming & Statistical Analysis
Vincent Lanzillo	Code Testing, Visualization & Final Report
Dean Dedios	Troubleshooting, Visualization & Report Research

API & Dataset Information: Kaggle Dataset Used:Tokyo 2020 Olympics | Kaggle

Two hundred six countries participated in the Tokyo games! We wanted to get more insight into the games, identify interesting stats and ideas. The dataset we used consisted of 11656 rows and 14 columns. We removed 159 rows because they had too many missing data fields. We ended up with a total of 11,497 rows and 14 columns. The dataset included 5 csv files that captured information about the athletes, coaches, medals won, and technical officials. The (5) csv files in the dataset were: Athletes.csv, coaches.csv, Medals.csv, medals_total.csv & technical_officials.csv.

The output tables and graphs are self explanatory but some interesting findings from the data analysis and statistical results included the following:

- The top 30 countries have 7931/11486 athletes, equivalent to 69% of total athletes.
- “Athletics” is the sport with the most participants (2033 people), “Swimming” came in second with 872 participants. Freestyle BMX cycling was the sport with the fewest participants (18 people).
- 93 countries earned at least one medal. The average medal count across the countries was 11.6. USA earned the most medals at 113 followed by China with 88. Japan had the highest gold medal count to total medals ratio *amongst the top 10* medal count leading countries with a value of .47.

Cleanup & Challenges

Cleaning the data was by far the most challenging aspect of the project. You will see different examples whereby we dealt with capitalizing first and last name [see 13] to removing missing data rows to changing column names [see 17] for standardization purposes. One hundred fifty-nine [see159] athletes did not have birth date, gender and discipline, [see 17], and so these rows were removed. We also noticed several columns including birth country, residence place, residence country and height ratio where these columns were not very interesting and significant for our analysis relatively speaking and so we dropped them as well, see [see 15].

Developing the graphs was also a bit of a challenge using matplotlib. As the Jupyter notebook shows, we mostly used pie and bar graphs. Displaying the graphs themselves was not as challenging. However, formatting the graphs to be aesthetically appealing was a bit challenging.

Statistical Analysis & Visualization

We developed some interesting ratios throughout the code looking at country gold and silver medal counts to total medal ratios for top countries. We also established various mean, standard deviation stats throughout [see 43,44].

The visualization examples are numerous and self-explanatory throughout the Jupyter notebook.

How else can you expand the project?

Our database analysis is for people that are interested in statistics for the Tokyo Olympic 2020 games -especially the relationship between athletes, country, gender, and medal. These elements in combination with disciplines (sport) produced a lot of interesting stats not only for entertainment purposes but also provided more insight on coaching/athlete training and competition research.

Our datasets were very straightforward and not very complex. As a means of expanding the project, it would be interesting if we could use the data to predict future Olympic performance, to validate or reject different what-if scenarios. Cultivating a series of hypotheses would give more purpose to the value of data, rather than appearing as random facts about our datasets. Additionally, to further improve our analyses, the group can add and append new and any missing data. As opposed to eliminating any parts of the datasets that contain a large amount of missing information, we can append any absent data by using other reliable sources to fill in the blanks. Little to no data removed allows for a much more reliable and accurate set of findings.

References Researched

[1] https://www.kaggle.com/piterfm/tokyo-2020-olympics?select=athletes.csv
[2] https://benalexkeen.com/bar-charts-in-matplotlib/
[3] https://www.pythoncharts.com/matplotlib/pie-chart-matplotlib/
[4] https://realpython.com/python-data-cleaning-numpy-pandas/
[5] Wes, McKinney. (2017). Python for Data Analysis (2nd ed.). O'Reilly Media.