*Farahin Choudhury*

*IST652: Scripting for Data Analysis*

***Python analysis: using pandas on olympics data***

# Introduction and Background

This past summer, the summer Olympics took place in Japan, with Tokyo being the host city of the games. The event drew a large audience (virtually) and resulted in many historic wins and losses. The objectives of this analysis is to use data from the Olympic games to analyze the findings from the following datasets provided in Excel files: athletes, coaches, gender, medals and teams. After importing all these files and converting them to data frames using the pandas module, the goal is to try and merge all these individual data frames into one master data frame, ideally called “OlympicsData”. If there are too many missing values generated from merging all the data frames, the code will be revisited and adjusted to merge data frames that make sense to be joined together for the purpose of the analysis.

# Data Source

The data was found on Kaggle, with all files created separately. The data sets can be found here: <https://www.kaggle.com/arjunprasadsarkhel/2021-olympics-in-tokyo>

Another data set was found on Kaggle in the JSON format, which can be found here: <https://www.kaggle.com/llui85/tokyo-2021-olympics-complete-grouped-by-type>

The JSON file is over 700 MB, but in an effort to add a layer of complexity to this analysis, this was attempted in the analysis (in a separate notebook file) to uncover any interesting findings that might pertain to this analysis. This ended up being quite difficult to parse through, as the keys resembled id’s which makes it difficult to differentiate the two; keys and id’s have different purposes in dictionaries.

# Data questions

There are several questions worth exploring throughout this analysis. Here are some questions of interest and relevance:

1. Which athlete received the highest number of medals?
   1. Which gender won the highest number of medals?
   2. Which were the top 5 sports for male and female athletes?
2. What is the average number of men and women that competed in the games?
3. Which countries among the biggest and smallest delegations had the highest medal count?
   1. Which countries won 50 or more medals?

# Python libraries used for this project

As the title of this report suggest, the pandas library is the main library behind this program. In addition to pandas, csv and numpy are the other important libraries to help read the Excel files as csv and to compute over to answer the data questions, respectively. Other libraries imported into the notebook were matplotlib, json and sklearn. Pandas proves to be quite powerful in manipulating the data frames to generate desired outputs.

# Data cleaning and preparation

This step of the analysis took the longest and was reviewed several times during the attempts to merge the data frames and answer the data questions. As the results of the data cleaning/preprocessing will show, it was not enough to convert the files into pandas data frames and to rename the columns. Several of the data frames had a column titled “NOC”, which stands for “Name of Olympic Committee”. This was instead renamed to “Olympic Committee Name”. The “Olympic Committee Name” is the official name for each delegation, while the “Country” is the specific country represented by the delegation. For example for the US, the country name is “United States” while the Olympic Committee name is the official name of the country “United States of America”.

Another common column encountered in various data frames is the “Discipline” column. This pertains to the actual sport the athletes competed in, so to make it more obvious, the column name was changed from “Discipline” to “Sport”.

To clean the data, the drop function came in handy. The ‘Rank’ column was dropped from the medals data frame, the NA values were dropped from the coaches data frame, and duplicate values found in the coaches were removed using the df.dropna() function. Without removing data with missing values, the data frame will look messy and calculations will be unsuccessful. This has the potential to generate inaccurate results compared to the true values of the outcome, but as previously stated the attempted calculations will generate errors.

The result of the data preparation for the teams data frame is displayed here:

A picture containing timeline

Description automatically generated

The code block where it reads **teams[‘Gender’].unique()** demonstrates the outcome of the effort after standardizing the values of ‘Gender’ to only ‘Men’, ‘Women, and ‘Other’.

# Results

This section of the report will show the results from the data cleaning and preprocessing stages and the output after merging the data frames, demonstrating the answers to the questions indicated in the “Data questions” section of the report. This section will also introduce new questions that were answered instead of the original questions listed above. While the code is explained in the notebook file, the output will also be shared in this report.

The main method of analysis in this project was creating data frames using the pandas module. The techniques that were used within the pandas framework included renaming columns, replacing values (such as the ‘Gender’ in the teams data frame), identifying unique values and dropping duplicate values entirely from the data frames. This allows for fine tuning the analysis and to focus on answering the data questions (and raising new questions along the way). Pandas also allows for parsing through the data frame using the .loc() and .groupby() methods.

The data frames that were merged to in order to attempt to answer the data questions were the total medals and teams data frames. The total medals data frame is actually a subset of the medals data frame, only including the names of the Olympic committees and the total medal count. If the medal types (gold, silver, bronze) were included, the merged data frame would generate many records with null values.

1. **Which gender won the highest number of medals?**

Graphical user interface, text, application

Description automatically generated

Even with ‘Other’ being a gender, this analysis reveals that women won the highest number of medals, followed by men (less than half than that of women) and then followed by ‘other’.

1. **What are the top and bottom 3 sports based on medal count?**

Table

Description automatically generated with medium confidence

From the merged mtdf data frame, this provides the whole list of values under “Total Medals”. Based on this result, the top three sports are **Swimming**, **Athletics** and **Fencing** and the bottom three sports are **Handball**, **Rhythmic** **Gymnastics** and **Baseball**/**Softball**.

1. **Which countries won 50 or more medals?**

**Table

Description automatically generated**

While not a result from a merged data frame, this is still an interesting finding from the Medals data frame. This screen shot captures the top 5 countries in descending order: **United States of America (US), People’ Republic of China (China), Japan, Great Britain and ROC (Russia**).

# Challenges

Most of the challenges were encountered during the ‘Data cleaning and preparation’ and ‘Results’ stages of the analysis. While the data types did not need changes (i.e. numbers didn’t need to be changed to numbers from strings), one of the trickier data frames to work with and standardize was the teams data frame. There were several different values under the “Gender” column, most of the values rendering them unnecessary; some examples are ‘4 x 400m Relay Mixed’, ‘Men's Team Pursuit’ and ‘Women's Épée Team’ (which is specific to only one sport in the entire data frame).

Also with the teams data frame, there were additional ‘Country’ values that were definitely not countries. The screen shot below shows the output for the different values that were contained in the ‘Country’ column:

Text

Description automatically generated

It is impossible to know (without individually searching via Google) which actual countries these athletes belonged in. Many of them, like *'Ludwig/Kozuch*' and *'Gaxiola/Rubio'*, appear as pairs but there is no information that can be found in any of the data frames that pertain to these pairs or teams. This is simply a data entry error on the data collector’s end.

# Conclusion/Learning Outcomes

Creating the data frames using pandas was a fun but challenging endeavor. Uncovering certain nuances like the different genders under the “Teams” data frame, required the code to be modified several times and different iterations of the code attempted. Even though the numeric data remained numeric and the strings were string objects, values of different data frames and the ways the data frames were merged needed to be refined. There are many code blocks with only one line of code to demonstrate the process of cleaning and preparing the data and to understand where there were issues with the data that would dictate the code to be changed.

When working through merging all the data frames together, the data frames generated values that were inconsistent with the original data they pulled from and did not yield the desired outcome. Merging two data frames separately fared a little better. However, it took several trials to know how to merge the data frames for this analysis.

The JSON module ended up not being utilized to further the analysis. An honest attempt was made but the data questions could not possibly be answered without accurately and successfully parsing through the JSON dictionaries. Perhaps this will be an asset for future projects after gaining better knowledge about JSON data but it was a good learning moment.

A screenshot of a computer

Description automatically generated with medium confidence

The output shown above is difficult to decipher. While it does show the values of the ‘Venue’ key from the olympicdata dictionary, there is no meaningful analysis or outcome that can be drawn from this. A greater expertise will allow for a more robust analysis.