# **Problem Statement**

The Olympic Games, considered to be the world's foremost sports competition has more than 200 nations participating across the Summer and Winter Games alternating by occurring every four years but two years apart.

In this project, I have taken the Olympics dataset(scraped from <https://en.wikipedia.org/wiki/All-time_Olympic_Games_medal_table),> to look at some interesting statistics and then try to find out which country is the King of the Olympic Games.

The dataset has details of 146 countries with the following 16 features

|  |  |
| --- | --- |
| **Feature** | **Description** |
| Country\_Name | Name of the country |
| # Summer | No. of games played in Summer Olympics |
| Gold\_Summer | No. of gold medals won in Summer Olympics |
| Silver\_Summer | No. of silver medals won in Summer Olympics |
| Bronze\_Summer | No. of bronze medals won in Summer Olympics |
| Total\_Summer | Total no. of all the medals won in Summer Olympics |
| # Winter | No. of games played in Winter Olympics |
| Gold\_Winter | No. of gold medals won in Winter Olympics |
| Silver\_Winter | No. of silver medals won in Winter Olympics |
| Bronze\_Winter | No. of bronze medals won in Winter Olympics |
| Total\_Winter | Total no. of all the medals won in Winter Olympics |
| # Games | Total no. of games played in both Summer and Winter Olympics |
| Gold\_Total | Total no. of gold medals won in both Summer and Winter Olympics |
| Silver\_Total | Total no. of silver medals won in both Summer and Winter Olympics |
| Bronze\_Total | Total no. of bronze medals won in both Summer and Winter Olympics |
| Total | Total no. of all the medals won in both Summer and Winter Olympics |

# **Motive**

In this project, I have applied the following concepts :

1. Dataframe operations
2. Conditional statement and loops
3. List operations
4. Bar Plotting
5. Mathematical operations

# **Step 1 - Data Loading**

* Load the dataframe from the path using pd.read\_csv() and store the dataframe in a variable called 'data'.
* In the dataframe, rename the column Total to Total\_Medals
* Display first 10 records using "head()" function to take a look at the dataframe.

# **Step 2 - Summer or Winter**

* Create a new column Better\_Event that stores 'Summer','Winter' or 'Both' based on the comparison between the total medals won in Summer event and Winter event (i.e. comparison between the Total\_Summer and Total\_Winter columns) using "np.where()"function.
* Find out which has been a better event with respect to all the performing countries by using value\_counts() function and store it in a new variable called 'better\_event'.

# **Step 3 - Top 10**

* Create a new dataframe subset called 'top\_countries' with the columns ['Country\_Name','Total\_Summer', 'Total\_Winter','Total\_Medals'] only
* Drop the last row from 'top\_countries'(The last row contains the sum of the medals)
* Create a function called 'top\_ten' that:
  + Takes the dataframe 'top\_countries' and a column name as parameters.
  + Creates a new empty list called 'country\_list'
  + Find the top 10 values for that particular column(for e.g. 'Total\_Summer') using "nlargest()" function
  + From the dataframe returned by nlargest function, slices the Country\_Name column and stores it in the 'country\_list' list
  + Returns the 'country\_list'
* Call the 'top\_ten()' function for the three columns :Total\_Summer,Total\_Winter and Total\_Medals and store their respective results in lists called 'top\_10\_summer', 'top\_10\_winter' and 'top\_10'
* Create a new list 'common' that stores the common elements between the three lists('top\_10\_summer', 'top\_10\_winter' and 'top\_10')

# **Step 4 - Plotting Top 10**

* Take the three previously created lists(top\_10\_summer, top\_10\_winter, top\_10)
* Subset the dataframe 'data' based on the country names present in the list top\_10\_summer using "isin()" function on the column Country\_Name. Store the new subsetted dataframes in 'summer\_df'. Do the similar operation using top\_10\_winter and top\_10 and store the subset dataframes in 'winter\_df' & 'top\_df' respectively.
* Take each subsetted dataframe and plot a bar graph between the country name and total medal count according to the event (For e.g. for 'summer\_df' plot a bar graph between Country\_Name and Total\_Summer)
* Modify the axes info accordingly.

# **Step 5 - Top performing country(Gold)**

* In the dataframe 'summer\_df'(created in the previous function) , create a new column Golden\_Ratio which is the quotient after dividing the two columns Gold\_Summer and Total\_Summer.
* Find the max value of Golden\_Ratio and the country associated with it and store them in summer\_max\_ratio and summer\_country\_gold respectively.
* In the dataframe 'winter\_df'(created in the previous function) , create a new column Golden\_Ratio which is the quotient after dividing the two columns Gold\_Winter and Total\_Winter.
* Find the max value of Golden\_Ratio and the country associated with it and store them in 'winter\_max\_ratio' and 'winter\_country\_gold' respectively.
* In the dataframe top\_df'(created in the previous function) , create a new column Golden\_Ratio which is the quotient after dividing the two columns Gold\_Total and Total\_Medals.
* Find the max value of Golden\_Ratio and the country associated with it and store them in top\_max\_ratio' and 'top\_country\_gold' respectively.

# **Step 6 - Best in the world**

* Drop the last row from the dataframe(The last row contains the total of all the values calculated vertically) and save the result in 'data\_1'
* Update the dataframe 'data\_1' to include a new column called Total\_Points which is a weighted value where each gold medal counts for 3 points, silver medals for 2 points, and bronze medals for 1 point. (i.e. You need to take the weighted value of Gold\_Total, Silver\_Total and Bronze\_Total)
* Find the max value of Total\_Points in 'data\_1' and the country associated with it and store it in variables 'most\_points' and 'best\_country' respectively.

# **Step 7 - Plot for the best**

* Create a single row dataframe called 'best' from 'data' where the value of column Country\_Name is equal to 'best\_country'(The variable you created in the previous task)
* Subset 'best' even further by only including the columns: ['Gold\_Total','Silver\_Total','Bronze\_Total']
* Create a stacked bar plot of 'best' using "DataFrame.plot.bar()" function
* Name the x-axis as United States using "plt.xlabel()"
* Name the y-axis as Medals Tally using "plt.ylabel()"
* Rotate the labels of x-axis by 45o using "plt.xticks()"