**Olympic Games – Dictionary**

Welcome to my project, this is my attempt at a Data Analysis project, I chose this field specifically just because I love watching the Olympics with all its different, cool and weird sports. It really has a lot to offer in form of entertainment and also a LOT of data that we can manipulate to our gain and if possible learn from them and keep improving.

I will touch on a few datasets here, that will allow me to get a broad idea on the visualizations I’m trying to reach. To be specific I will mention every column and what it gives, these datasets include:

* athletes\_events.csv: this one has data on a massive number of athletes from 1896 to 2016, the columns that exist here are all related to each athlete’s info like: physique, country, medals.
* ID: Number that is specific to each athlete
* Name: Full name of the athlete, including their middle names and other additions
* Sex: Female or Male
* Age: Age at the time of competing in that year
* Height: One uniform height measurement
* Weight: One uniform weight measurement
* Team: Country represented by the Athlete
* NOC: Code of that Country
* Games: Combination of the Year and Season in which that athlete played
* Year: The year the athlete participated in
* Season: Winter or Summer games
* City: The city in which the games were held
* Sport: The general name of the sport that the athlete participated in
* Event: Distinct type of the sport
* Medal: Type of Medal the athlete won, Gold, Silver, Bronze or NaN

I stripped a lot of data from this dataset as I didn’t need the rows that had anything to do with the Winter games, I only kept the Summer data, and by that I had to drop the “Games” and “Season” columns because they gave me no additional info about the athlete or the event

I named the Summer equivalent of this dataset ‘athlete\_events\_s’, the suffix ‘\_s’ refers to summer games and subsequently the ‘\_t’ refers to the total of both winter and summer.

Source : https://www.kaggle.com/marcogdepinto/let-s-discover-more-about-the-olympic-games/data

* olympics.csv: this dataset has data about countries that participated in the Olympics. It includes info like the medals won, first appearance, …

All the info about the columns are provided in the link below but as before I have to get rid of data about the winter game so I’ll drop of a few rows and columns: "Medal", "W\_Medal", "Apps", "Medal.1", "Gold", "Silver", "Bronze", 'WO\_Apps', 'WO\_Medal', 'WO\_Gold', 'WO\_Silver', 'WO\_Bronze'

And just like the previous dataset I will be naming adding the ‘\_s’ suffix to the summer related data.

Source : <https://data.world/johayes13/summer-winter-olympic-games/workspace/data-dictionary>

The next three datasets have info about the 2021 Tokyo Olympics

* tokyo\_athletes :
* Name: Name of the Athlete
* NOC: Nationality of that athlete
* Discipline: The event in which the athlete participated in
* tokyo\_genders :
* Discipline: The events that help place
* Female: Number of female athletes
* Male: Number of male athletes
* Total: Total of athletes that participated in the event
* tokyo\_medals :
* Rank: Rank of the country by number of gold medals won
* Team/NOC: Name of the country
* Gold: Number of Gold won by each country
* Silver: Number of Silver won by each country
* Bronze: Number of Bronze won by each country
* Total: Total number of medals won by each country
* Rank by Total: Rank of the country by total number of medals won
* noc: This DataFrame will help me later in reorganizing and building new dataframes based on the country codes
* Team/NOC: Country name
* Code: Country Code

⚠There will be some minor dataframes that I will not mention as they’re there as a bridge to help me reach another point of the build but won’t play a role in later visualizations

The square bullet below is used only to showcase the dataframes I’ll use to build visualizations

* First DataFrame: it’s called ‘updated’ and it sums up both info about the medals won in the ‘olympics\_s’ and ‘tokyo\_medals’, so we’re kinda updating the first df, the updated df will have the same columns as the first one but with more count of medals if the country has won any
* Second DataFrame: ‘arab\_olympics’ which is the same as before just filtered out the Arab countries
* Third DataFrame: ‘noc\_genders’ is a table that has info about the gender of athletes that participated in the Olympics for each country, its columns are:
  + Country: Name of the Country
  + Code: Code of that country
  + Male: Sum of male athletes who represented that country
  + Female: Sum of female athletes who represented that country
  + Total: Sum of all athletes
* Fourth & Fifth DataFrames: we have here two DFs ‘evnt\_genders’ and ‘all\_genders’, they both represent the same data and it’s the count of female and male athletes per Discipline, but the first has info just about the ‘athlete\_events\_s’ DF and the second merges data from ‘athlete\_events\_s’ and ‘tokyo\_medals’ so it kinda updates the count to 2021, their columns are:
  + Discipline: The Sport which those athletes played
  + Male: Sum of male athletes who played it
  + Female: Sum of female athletes who played it
  + Total: Sum of all athletes
* Sixth & Seventh DataFrames: ‘succ\_athletes’ which includes just the athletes that were successful in the Olympics by winning any kind of Medal, and it has the same columns as before, and then comes ‘succ\_ath’ that has info about each athlete and the number of medals won, its columns are:
  + Name: Name of the athlete
  + NOC: Code of the country the athlete represents
  + Apps: Number of times the athlete participated in the Olympics
  + Gold: Sum of gold medals won
  + Silver: Sum of silver medals won
  + Bronze: Sum of bronze medals won
  + Total: Sum of all Medals won
  + Ratio(T/A): Ratio of total medals won per participations
  + Sport: Specialty sport that the athlete won the medals in
* Eight DataFrame: ‘age\_event’ this one takes a few popular sports and gives us a sneak peak of the age difference between them all and through the years, this is done by calculating the mean age of the athletes by each sport. A few info about some sports were missing so I replaced them with the average. Its columns are:
  + Year: Period in which the average was calculated
  + Gymnastics
  + Swimming
  + Athletics
  + Football
  + Weightlifting
  + Basketball
  + Volleyball
  + Water Sport
  + Tennis
  + Rugby
* Ninth & Tenth DataFrames: ‘physique\_h’, ‘physique\_w’ both are snippets from the ‘succ\_ath’ but I kept the columns concerning the physique info of the athletes and then I dropped the rows that were missing Height or Weight info for each DF, the columns are the same except for the first I kept just Height and the second just Weight