Dataset - 120 Years of Olympic History Athletes and Results

Steps

- Saved both files as an Excel Worksheet from CSV format.
- Removed duplicates across the whole row values.
- Removed ‘notes’ column from ‘noc_regions’ table.
- Replaced “NA” with null from the ‘Age’, ‘Height’ and ‘Weight’ columns.
- Replaced “NA” with "No Medal" in the ‘Medal’ column.
- Removed ‘Games’ column, because the year and season info is available in two individual columns.
- Created a ‘Sport’ wise grouping table with averages of ‘Age’, ‘Height’, ‘Weight’. Due to different ‘Sport’ having different ages, heights, and weights, filling global average would be inaccurate, so I categorise the average by the sport.
- Merged the grouping table with the main table using merge queries and expanded with the 3 average columns.
- Added 3 custom columns with a condition that if the main column values are null, then place the average value I got from the grouped table. Some ‘Sport’ don’t have average values, for that case I filled the global average of the respective column.
- Deleted the old columns and the other average columns taken from the grouped table.
- In the ‘Name’ column, some values contain nicknames in quotes and alternate names in parentheses, so removed them using the find and replace option. For values in double quotes, used “*” and values in brackets, used (*).
- Added a new column and used the Trim function to remove extra spaces added when doing the last step. Using paste special, pasted the formula into values and deleted the old ‘Name’ column.
- Replaced “NA” with “Unknown” in the ‘noc_regions’ sheet on the ‘Region’ column.