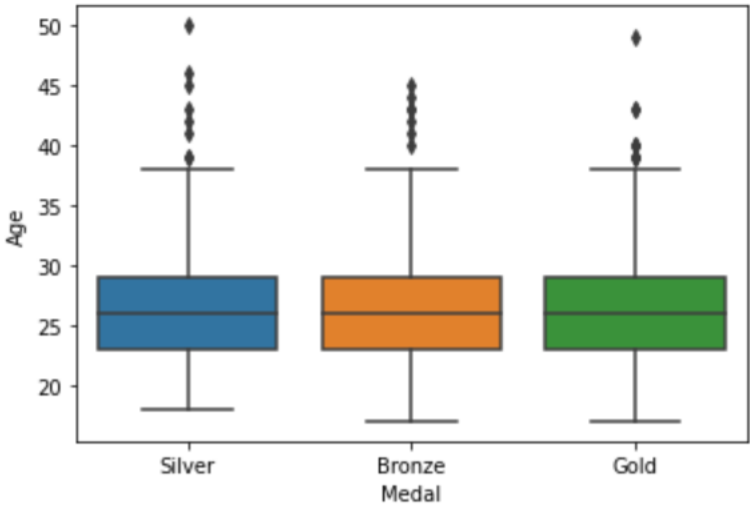
REPORT

Data cleaning :

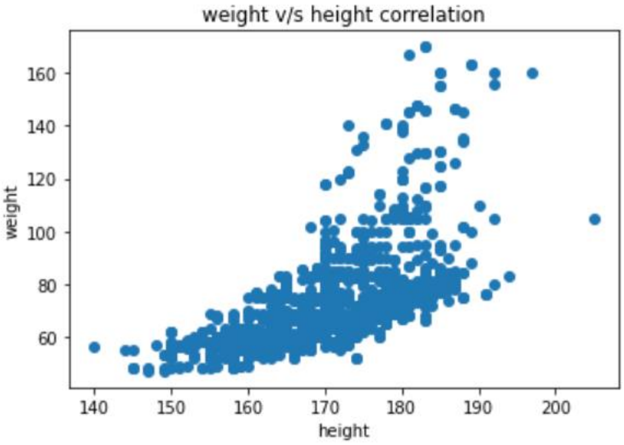
- Checking for null data revealed that only the age, height & weight columns had empty cells.
- As all three are numerical values, these empty cells were filled in with the mean from their respective columns.
- Upon further inspection (as will be elaborated in the data visualization component), it was revealed that only 2.37% of the numerical age data, 0.09% of the numerical height data and 5.61% of the numerical weight data were outliers and hence, it was deemed that outlier deletion was unnecessary

Data visualization :

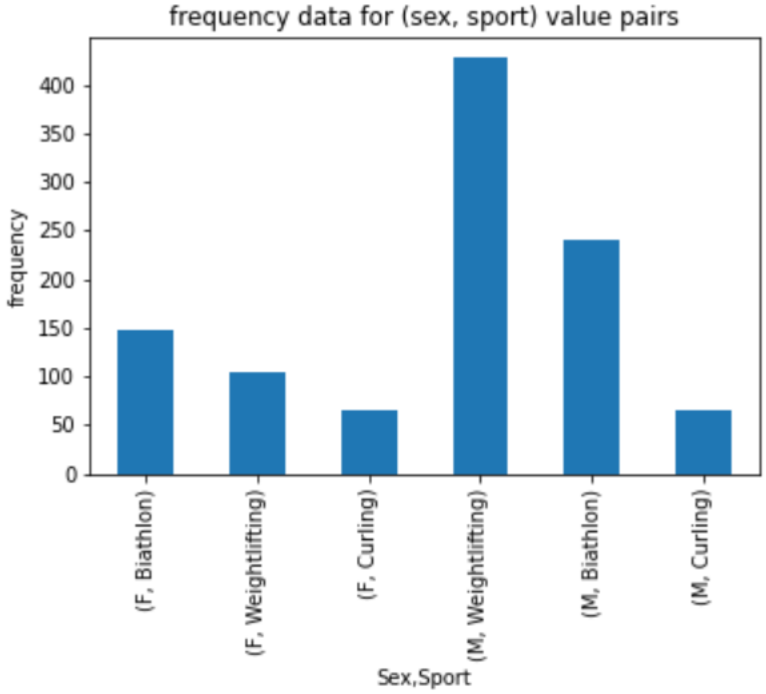
1. The age distribution of silver medallists turned out as shown in the attached boxplot -



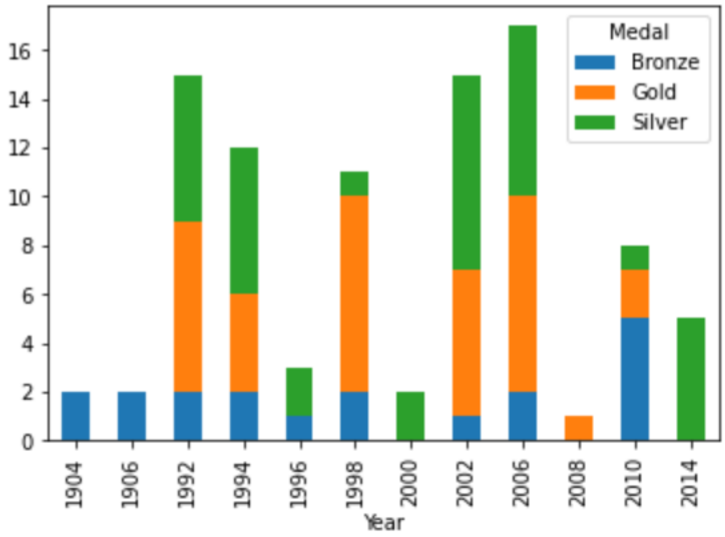
2. <Outlier analysis was done at this point, as elaborated before>
3. As per introductory question 4, a scatterplot of the athletes' height/weight distribution was made as follows and it was inferred that there was a positive exponential correlation between the two, highly suitable for use with a polynomial regression model should the data be used as such.



4. As per introductory question 5, manual evaluation of the dataset indicated that there were no listed sports (of the three given i.e.; biathlon, weightlifting & curling) that had less than 5 years of data. This was determined by taking the difference of the minimum year and the maximum year for each sport.
5. The gender distribution for each of the aforementioned sports turned out as follows –



6. As per task question 1, it was found (as demonstrated in the attached python notebook) that Germany won the highest number of medals all in all.
7. Carrying forward from that inference, the country's distribution of medals won could be represented as follows –



Dataset & attribute creation :

- As per introductory question 3, a new column was appended into the existing dataset containing the BMI of each of the listed athletes.

- Similarly, a new dataframe was created as per task question 2 where the corresponding number of medals was divided by the number of years of participation by each athlete to get a measure of their success. This was extracted as the attached CSV file (task2_result.csv)