Introduction

The dataset contains more than 35,000 rows about athletes who participated in the Tokyo Olympics 2020. The point of focus for this project is to see how Team USA fared at this global event.

Data Quality Issues

Right off the bat, I observed data redundancy arising from column *"Team_Members_Select"*, so I deleted the column to clean up the duplicates. This led the original dataset to shrink from 38,000 rows to 24,000.

The dataset also had a column "*Person_Age_Days*" to denote the age of participants, but it also had inconsistencies arising due to null values and multiple values arising because of different event dates . As a result, a new column was created to indicate the age in years.

Additionally, I observed that the *"Competitor"* column had instances where the person ID was -1 which reflected a row representing a team event.

Analysis

The USA has consistently been a top performing country in olympics. In order to compare the performance of the USA with the top- 5 countries in the final medal tally, I first analysed the dataset to realise that the dataset contained a list of individual participants and thus, each member of a team event will have the same final rank. To deal with this, I grouped rows on the basis of Event_ID and Team_ID.This helped me create a visualization with top-5 countries in the medal count and compare the performance of the US with other countries. I represented this using a bar chart with filters for Gold, Silver, Bronze, and Total medals.

I was interested to see the sports in which USA participants ended up at position 4. This is insightful because it helps us see which sports had athletes very close to winning a medal but failed to cross the line because of some reason. This can help us visualise the list of sports where team USA can pay more attention and make sure the line is crossed in 2024.

Additionally, I wanted to identify individual athletes from Team USA who secured the most medals and highlight standout performers and their respective events.

To break down performance by specific sports to see which events Team USA excelled in, I created a bubble chart visualisation.

In the end, to see how the ages of men and women were distributed across participants, I created a butterfly chart.

For additional analysis, please refer to the presentation slides uploaded along with this email.