Project: Apply data wrangling to your capstone project

To clean my data, I started with checking the structure. Then considering I wanted to evaluate the age, sport and team components, I checked to see which columns had missing values. Age was one of the columns with missing values so first I checked updating missing values with the mean age, however checking via a few sports, I realized this could skew the results so I decided to remove all rows with NA for age. I do believe removing these rows will provide better results as the mean results could misrepresent the age of those athletes. I also mutated those blank under Medals to state “no medal earned” I have also added columns for Age/Sport and Sport/Year to allow to get counts for frequency. I have run a table to check for low Sport events and filtered to keep those sports with over 4000 data entries, leaving 20 Sports. I followed similar wrangling of the Team data, keeping those teams with over 20 entries, unfortunately this reduced data to far.