Daniel Yi Hong

Step 1.

What are the preferred/ideal age, height and weight in each sport separated by genders?

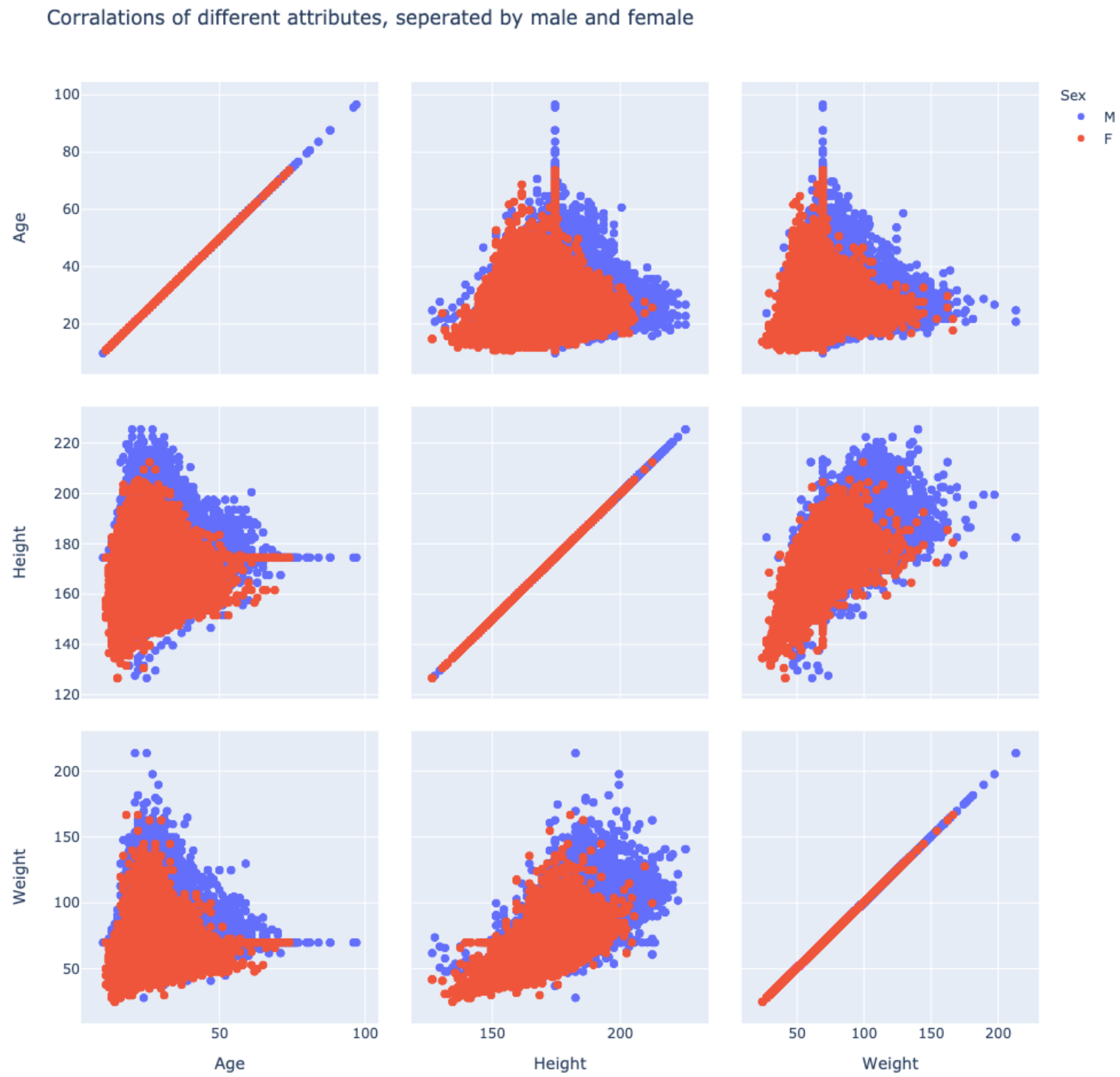
Step 2.

Upon inspection of the data set, I observe that the data cannot be used directly for plotting without further processing, there are entries (age, height, weight) that have missing numerical values and filled with **'NA'.**

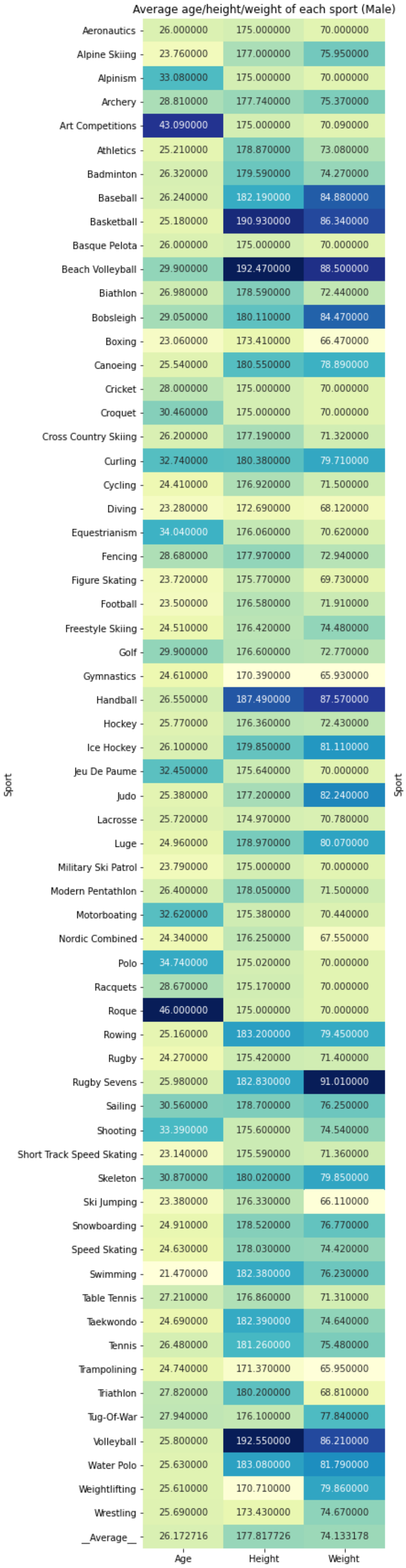
Therefore, I decide to fill entries that have missing numerical values with the mean value of corresponding columns. I also use several other data processing techniques, such as pivot, to pivot the 'Sport' column to index column, and groupby, to group entries that have the same index together. Min-Max scaling method is also being applied for some graphs. I use both Plotly and Seaborn library to complete the visualizations.

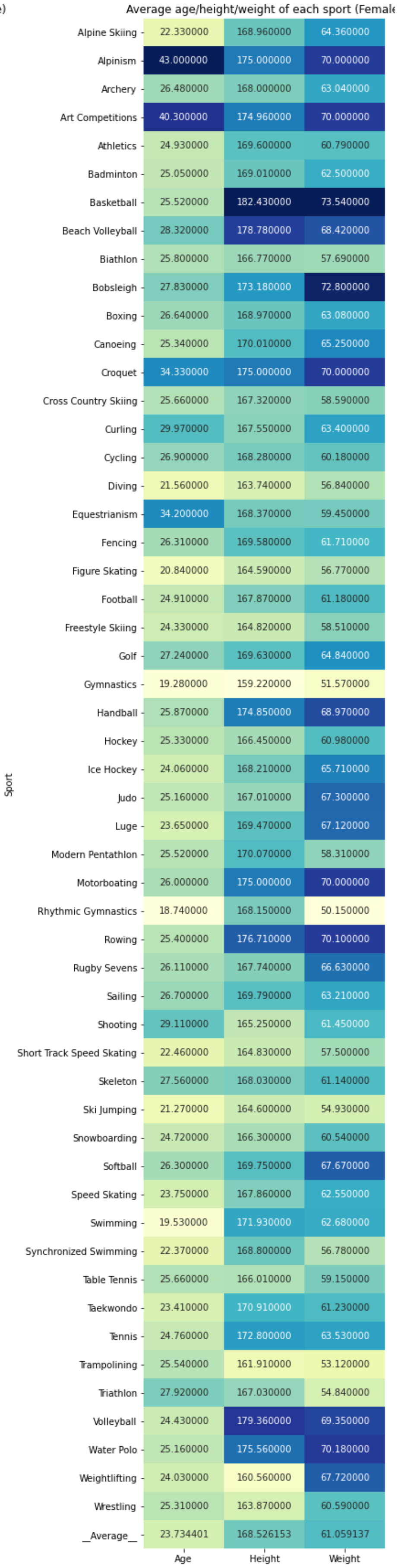
Step 3.

First, in order to get a clearer view of the correlations between the target attributes, I plot the correlations in scatter plot, use different colours to separate genders:



From the plot, we can see that height and weight are positively correlated; at the same height, males are heavier than females. Age and height/weight don't show much of correlation.
Next, I plot a heatmap that has age, height, weight as x-axis and sport as y-axis, separated by genders:

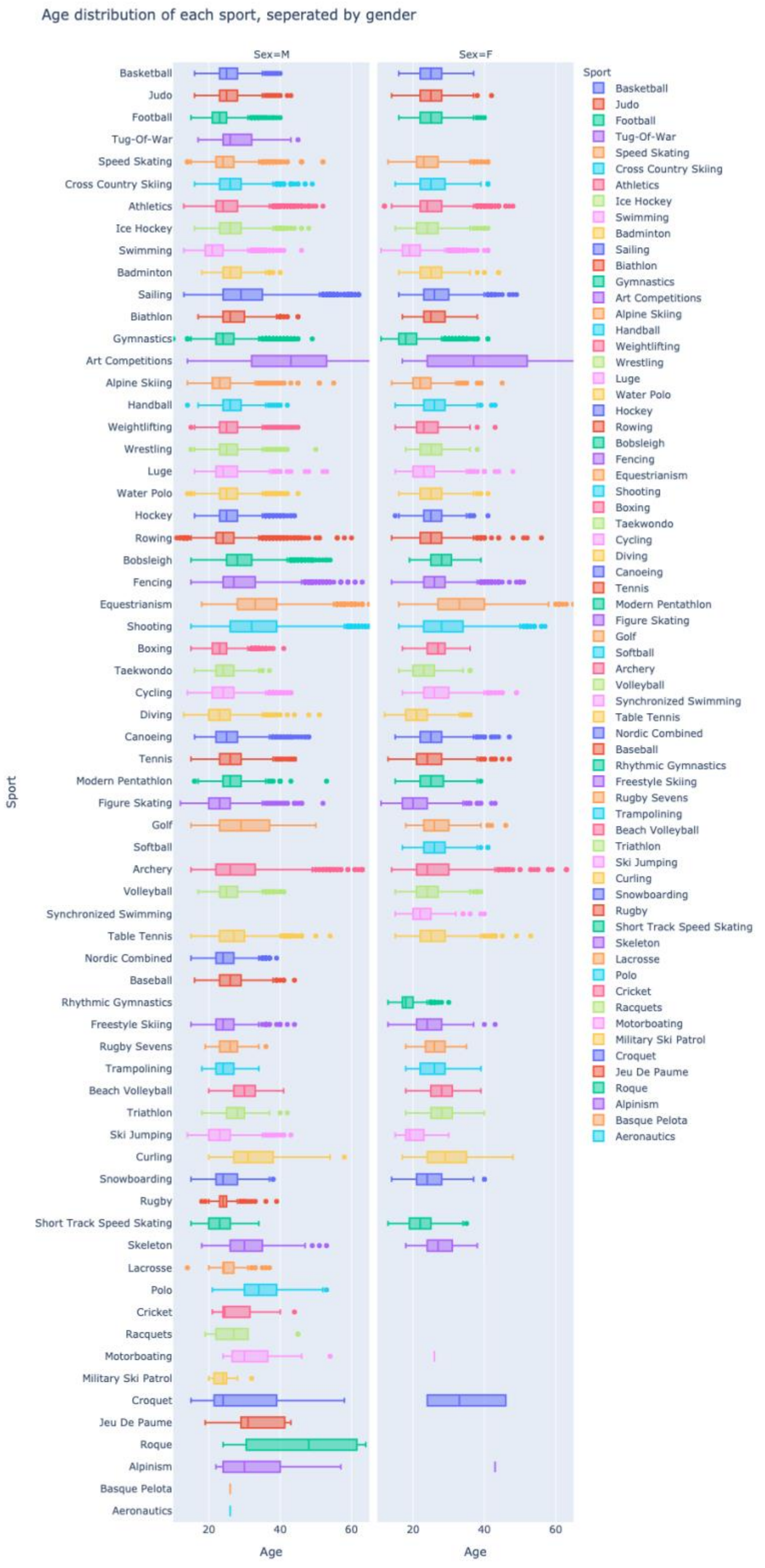


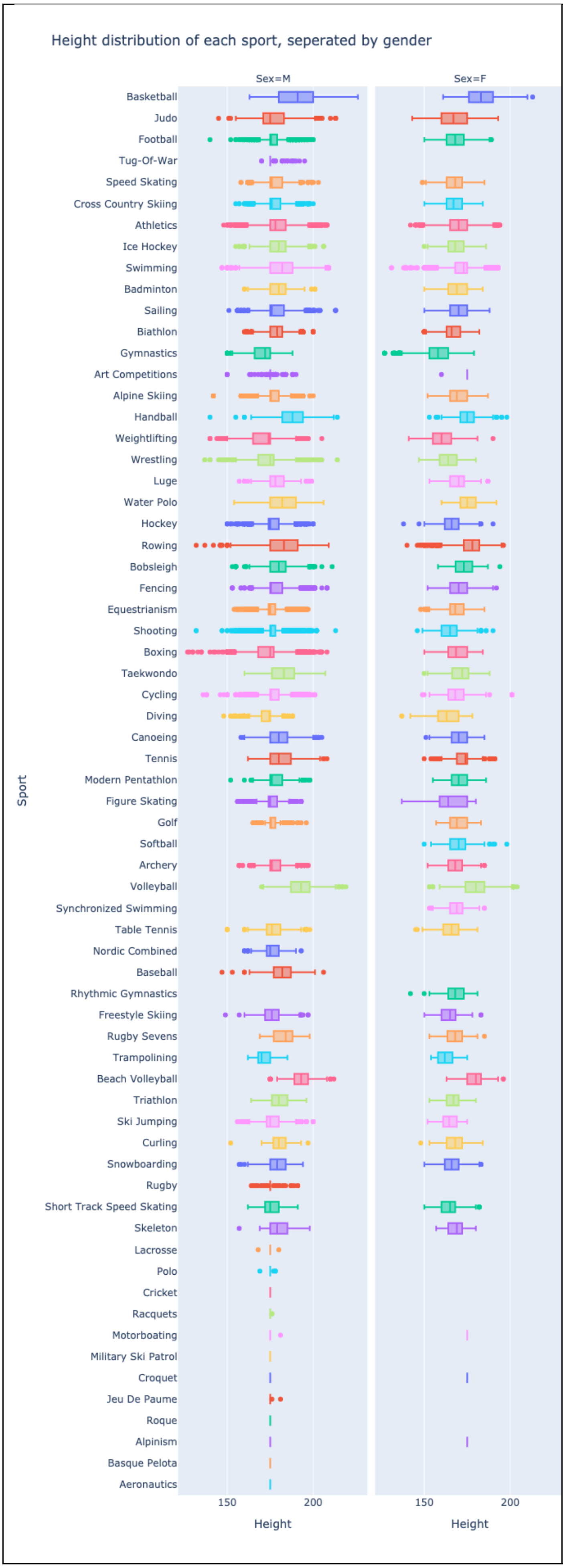


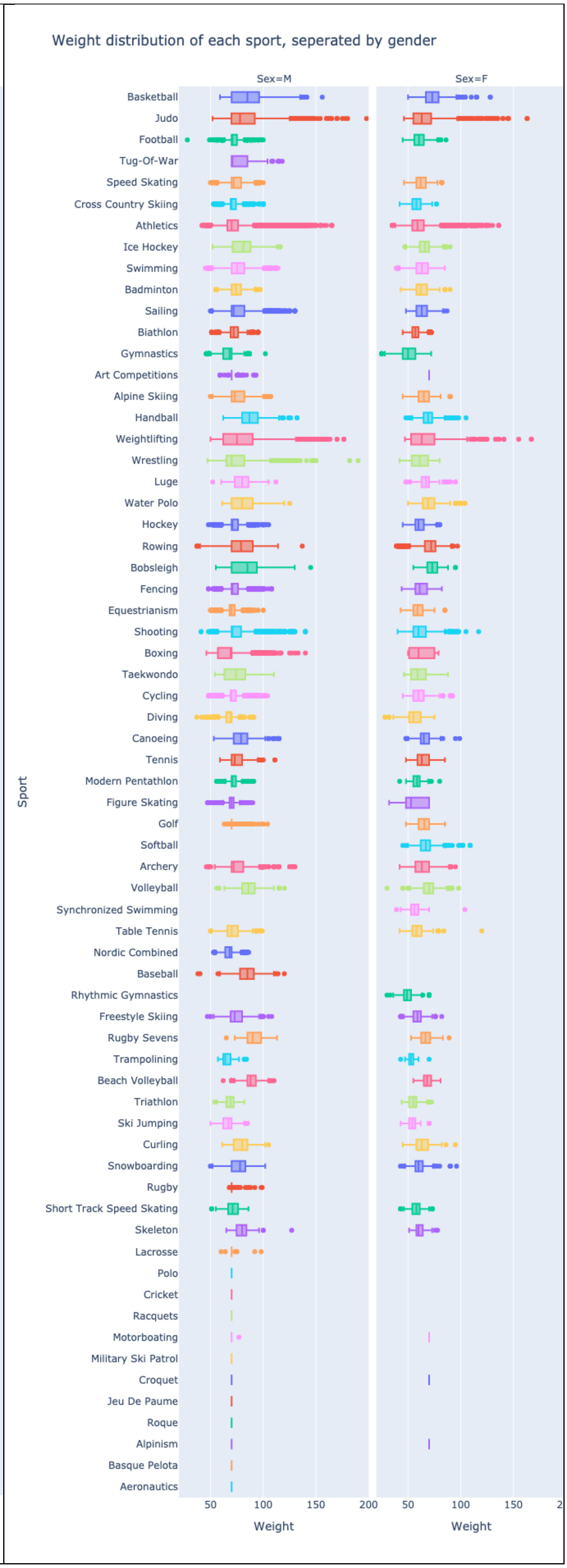
The plot shows the following information:

- For “Age” attribute, in the male group, Swimming and Boxing have the lowest mean age. Roque and Art competitions have highest mean age. In female group, Swimming and Gymnastics have lowest mean age while Alpinism and Art competitions have the highest.
- For “Height” attribute, in the male group, Basketball and Volleyball have the largest mean height while Gymnastic and Weightlifting have the lowest. In the female group, it’s exactly the same.
- For “Weight” attribute, in the male group, Gymnastics and Trampolining have the lowest weight while Beach volleyball and Rugby have the highest. In the female group, Basketball and Bobsleigh have the largest weight while Gymnastics and Trampolining have the lowest weight.

Next I use box plots. I use Age/Weight/Height as x-axis, Sport as y-axis, and box for distributions:




Height distribution of each sport, seperated by gender
Sex=M
Sex=F
Sport
Basketball
Judo
Football
Tug-Of-War
Speed Skating
Cross Country Skiing
Athletics
Ice Hockey
Swimming
Badminton
Sailing
Biathlon
Gymnastics
Art Competitions
Alpine Skiing
Handball
Weightlifting
Wrestling
Luge
Water Polo
Hockey
Rowing
Bobsleigh
Fencing
Equestrianism
Shooting
Boxing
Taekwondo
Cycling
Diving
Canoeing
Tennis
Modern Pentathlon
Figure Skating
Golf
Softball
Archery
Volleyball
Synchronized Swimming
Table Tennis
Nordic Combined
Baseball
Rhythmic Gymnastics
Freestyle Skiing
Rugby Sevens
Trampolining
Beach Volleyball
Triathlon
Ski Jumping
Curling
Snowboarding
Rugby
Short Track Speed Skating
Skeleton
Lacrosse
Polo
Cricket
Racquets
Motorboating
Military Ski Patrol
Croquet
Jeu De Paume
Roque
Alpinism
Basque Pelota
Aeronautics
150
200
150
200
Height
Height


Weight distribution of each sport, seperated by gender
Sex=M
Sex=F
Sport
Basketball
Judo
Football
Tug-Of-War
Speed Skating
Cross Country Skiing
Athletics
Ice Hockey
Swimming
Badminton
Sailing
Biathlon
Gymnastics
Art Competitions
Alpine Skiing
Handball
Weightlifting
Wrestling
Luge
Water Polo
Hockey
Rowing
Bobsleigh
Fencing
Equestrianism
Shooting
Boxing
Taekwondo
Cycling
Diving
Canoeing
Tennis
Modern Pentathlon
Figure Skating
Golf
Softball
Archery
Volleyball
Synchronized Swimming
Table Tennis
Nordic Combined
Baseball
Rhythmic Gymnastics
Freestyle Skiing
Rugby Sevens
Trampolining
Beach Volleyball
Triathlon
Ski Jumping
Curling
Snowboarding
Rugby
Short Track Speed Skating
Skeleton
Lacrosse
Polo
Cricket
Racquets
Motorboating
Military Ski Patrol
Croquet
Jeu De Paume
Roque
Alpinism
Basque Pelota
Aeronautics
50
100
150
200
50
100
150
Weight
Weight

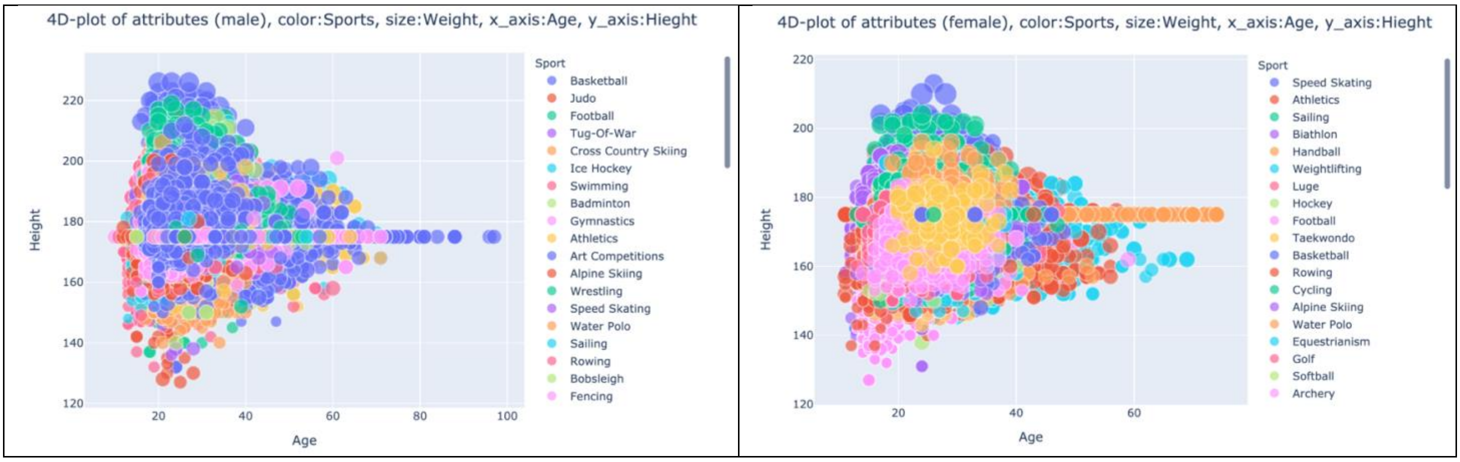
The graph provides the same information as the heatmap but in more details:

- The distributions of each attribute are shown instead of just a mean number. In .py file, when users hover over the box in the plot, min/max, mean and bounds are shown.
- Since some sport are male only and others are female only, heatmap does not align all the same sports at a same horizontal line. Box plot solved this issue allow an easier comparison.
- However, given the nature of the data, box plot is not as simplistic as heatmap.

After that, I plot a multi pie chart, each sport is a pie, and attributes in the pie shows their contribution to each sport, i.e., less Age% means younger age in the particular sport, more weight% means higher weight.

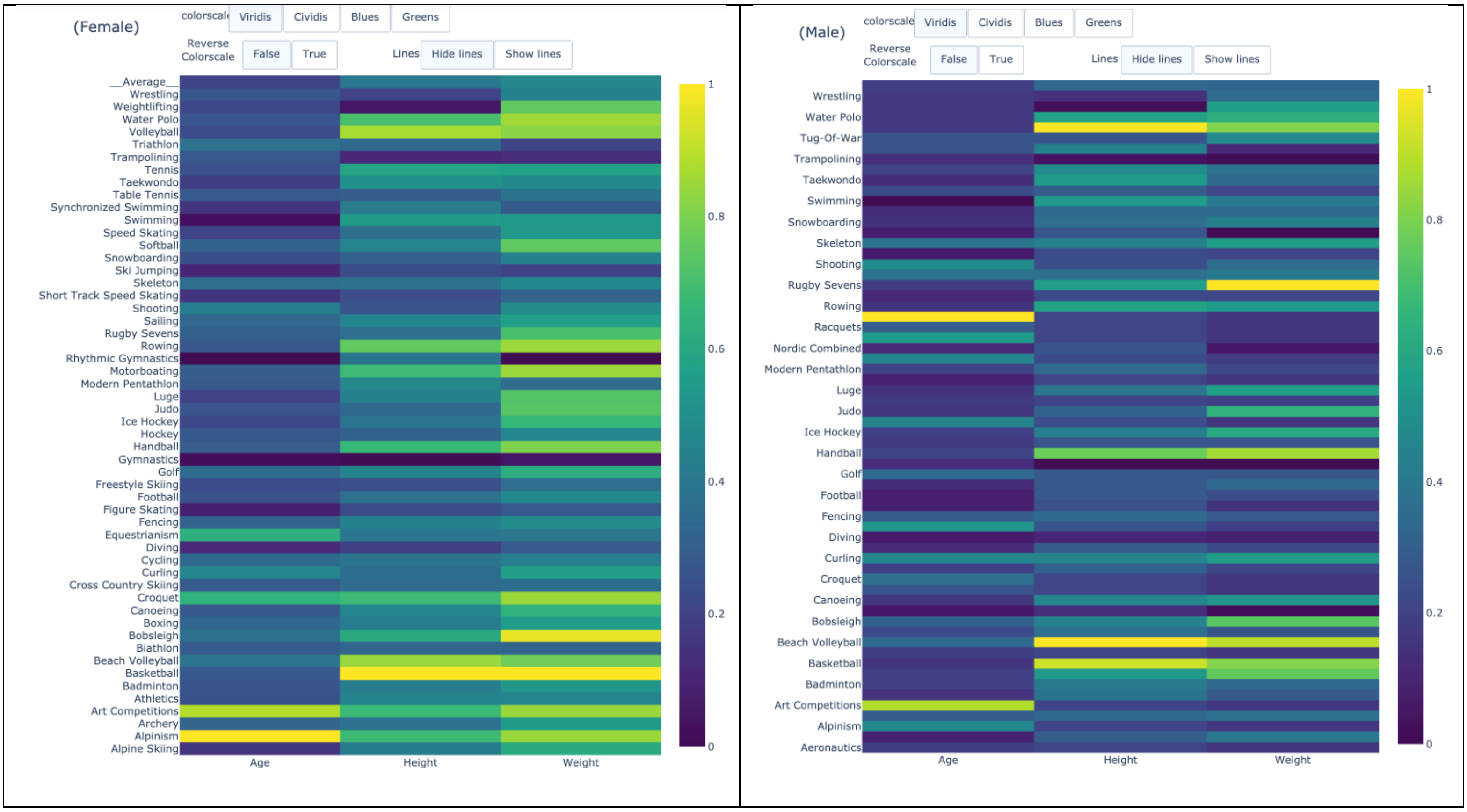
We can see that it is difficult to clearly compare each attribute between male and female, and it is hard to navigate between sports. However, it is easier to see how each attribute contribute to each sport.

The next is a 4-D plot to visualize the correlation between sports and each attributes.



We can see in this type of graphs is more suitable user interactions, such as users can hover mouse over each circle and discover information that pops up and zoom in and zoom out to get better views. But as a standstill image, it does not have advantage over other plots.

For interactive design, I also plot a heatmap that allows user to control colours and lines of the heatmap:



We can see that while allowing users to control colours is great, however, in this dataset, given the users the ability to see numbers directly is more helpful, since there are too many categories of sports.

Analysis to answer the question in step 1:

- Graphs show that different sports can be bias towards different attributes or different combination of attributes. For the same sport, the bias applies to both genders the same way.
- If I were to choose a best sport for me, I would select top 10 sports in each attribute (30 total) that are best for me, then find an intersection of these, then pick one of them based on my interest.
- To choose a best visualization technique, I would pick the box plot, since it provides the best detailed information and allow user interaction (on .py file). Comparisons are clear between genders. Views are clear, and I can quickly find information from each attribute without much efforts.

Overall, the information obtained from the graphs coincides with common sense. It proves that the analysing techniques are effective.

EXTRA INFO:

(Note: I decided not to investigate 'Medal' information since each country entered Olympic in different years, and it is impossible to determine which year as starting point, hence medal count can be biased.)

I plot the athletes' age/height/weight changes over the years. It is quite interesting.