## Step 1.

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

## Step 2.

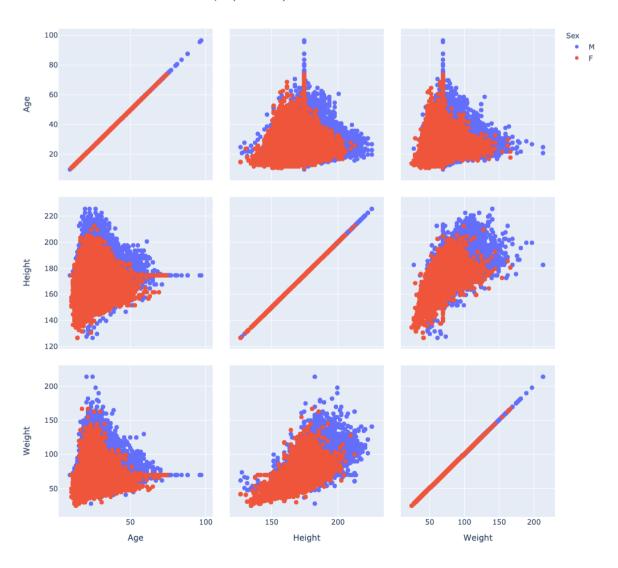
Upon inspection of the data set, I observe that the data cannot be used directly for plotting without further processing, there are entries (age, height, weight) that have missing numerical values and filled with 'NA'.

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:

Corralations of different attributes, seperated by male and female



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:

A	verage age/hei	ight/weight of e	each sport (Mal	e) Ave	rage age/heig	ht/weight of ea	ch sport (Female
Aeronautics -	26.000000	175.000000	70.000000	Alpine Skiing -	22.330000	168.960000	64.360000
Alpine Skiing -		177.000000	75.950000	Alpinism -	43.000000		70.000000
Alpinism -		175.000000	70.000000	Archery -	26.480000	168.000000	63.040000
Archery -	28.810000	177.740000	75.370000	Art Competitions -	40.300000	174.960000	70.000000
Art Competitions -		175.000000 178.870000	70.090000	Athletics -	24.930000	169.600000	60.790000
Athletics - Badminton -		179.590000	74.270000	Badminton -	25.050000	169.010000	62.500000
Baseball -		182.190000	84.880000				
Basketball -		190.930000	86.340000	Basketball -	25.520000	182.430000	73.540000
Basque Pelota -		175.000000	70.000000	Beach Volleyball -	28.320000	178.780000	68.420000
Beach Volleyball -		192.470000	88.500000	Biathlon -	25.800000	166.770000	57.690000
Biathlon -	26.980000	178.590000	72.440000	Bobsleigh -	27.830000		72.800000
Bobsleigh -	29.050000	180.110000	84.470000	Boxing -	26.640000	168.970000	63.080000
Boxing -	23.060000	173.410000	66.470000	Canoeing -	25.340000	170.010000	65.250000
Canoeing -	25.540000	180.550000	78.890000	Croquet -	34.330000	175.000000	70.000000
Cricket -	- 28.000000	175.000000	70.000000	Cross Country Skiing -	25.660000	167.320000	58.590000
Croquet -	30.460000	175.000000	70.000000		29.970000	167.550000	63.400000
Cross Country Skiing -	26.200000	177.190000	71.320000	Curling -			
Curling -	32.740000	180.380000	79.710000	Cycling -	26.900000	168.280000	60.180000
Cycling -		176.920000	71.500000	Diving -	21.560000	163.740000	56.840000
Diving -		172.690000	68.120000	Equestrianism -	34.200000	168.370000	59.450000
Equestrianism -	- 34.040000 - 28.680000	176.060000 177.970000	70.620000 72.940000	Fencing -	26.310000	169.580000	61.710000
Fencing - Figure Skating -		177.970000	69.730000	Figure Skating -	20.840000	164.590000	56.770000
Football -		176.580000	71.910000	Football -	24.910000	167.870000	61.180000
Freestyle Skiing -		176.420000	74.480000	Freestyle Skiing -	24.330000	164.820000	58.510000
Golf -		176.600000	72.770000	Golf -	27.240000	169.630000	64.840000
Gymnastics -	24.610000	170.390000	65.930000			159.220000	51.570000
Handball -	26.550000	187.490000	87.570000	Gymnastics -			
Hockey -	25.770000	176.360000	72.430000	Handball -	25.870000	174.850000	68.970000
Ice Hockey -	26.100000	179.850000	81.110000	Hockey -	25.330000	166.450000	60.980000
Jeu De Paume -	32.450000	175.640000	70.000000	Ice Hockey -	24.060000	168.210000	65.710000
jado		177.200000	82.240000	ॐ Judo -	25.160000	167.010000	67.300000
Lacrosse -		174.970000	70.780000	Luge -	23.650000	169.470000	67.120000
Luge -		178.970000	80.070000	Modern Pentathion -	25.520000	170.070000	58.310000
Military Ski Patrol -		175.000000 178.050000	70.000000	Motorboating -	26.000000	175.000000	70.000000
Modern Pentathlon - Motorboating -		175.380000	70.440000	Rhythmic Gymnastics -	18.740000	168.150000	50.150000
Nordic Combined -		176.250000	67.550000	Rowing -	25.400000	176.710000	70.100000
Polo -		175.020000	70.000000		26.110000	167.740000	66.630000
Racquets -		175.170000	70.000000	Rugby Sevens -			
Roque -	46.000000	175.000000	70.000000	Sailing -	26.700000	169.790000	63.210000
Rowing -	25.160000	183.200000	79.450000	Shooting -	29.110000	165.250000	61.450000
Rugby -	24.270000	175.420000	71.400000	Short Track Speed Skating -	22.460000	164.830000	57.500000
Rugby Sevens -	25.980000	182.830000	91.010000	Skeleton -	27.560000	168.030000	61.140000
Sailing -	30.560000	178.700000	76.250000	Ski Jumping -	21.270000	164.600000	54.930000
Shooting -	33.390000	175.600000	74.540000	Snowboarding -	24.720000	166.300000	60.540000
Short Track Speed Skating -		175.590000	71.360000	Softball -	26.300000	169.750000	67.670000
Skeleton -	30.870000	180.020000	79.850000	Speed Skating -	23.750000	167.860000	62.550000
Ski Jumping - Snowboarding -		176.330000 178.520000	66.110000 76.770000	Swimming -	19.530000	171.930000	62.680000
Speed Skating -		178.030000	74.420000	Synchronized Swimming -	22.370000	168.800000	56.780000
Swimming -		182.380000	76.230000			166.010000	
Table Tennis -		176.860000	71.310000	Table Tennis -	25.660000		59.150000
Taekwondo -		182.390000	74.640000	Taekwondo -	23.410000		61.230000
Tennis -	- 26.480000	181.260000	75.480000	Tennis -	24.760000	172.800000	63.530000
Trampolining -	24.740000	171.370000	65.950000	Trampolining -	25.540000	161.910000	53.120000
Triathlon -	- 27.820000	180.200000	68.810000	Triathlon -	27.920000	167.030000	54.840000
Tug-Of-War	27.940000	176.100000	77.840000	Volleyball -	24.430000	179.360000	69.350000
Volleyball -		192.550000	86.210000	Water Polo -	25.160000	175.560000	70.180000
Water Polo -		183.080000	81.790000	Weightlifting -	24.030000	160.560000	67.720000
Weightlifting -	25.610000	170.710000	79.860000	Wrestling -	25.310000	163.870000	60.590000
Wrestling -		173.430000	74.670000		23.734401	168.526153	61.059137
_Average	- 26.172716 Age	177.817726 Height	74.133178 Weight	_Average	23.734401 Age	Height	Weight
	нус	rieigilt	vicignit		Aye	rieigilt	vicigiit

The plot shows the following information:

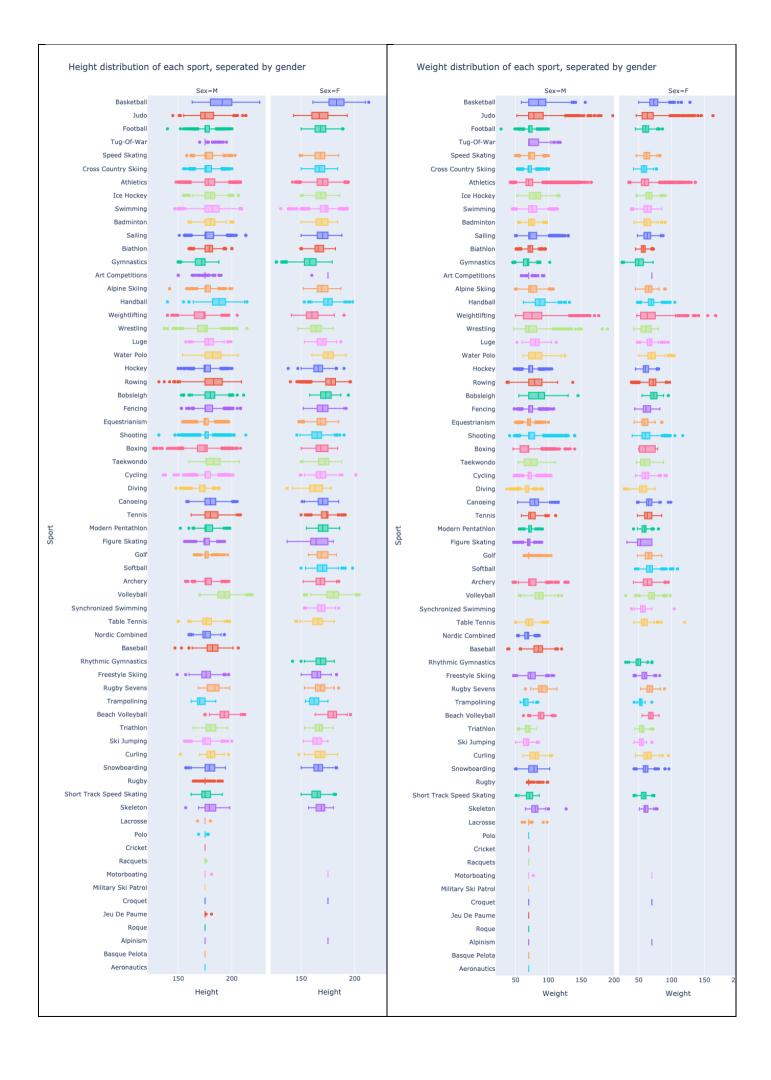
- For "Age" attribute, in the male group, Swimming and Boxing have the lowest mean age. Roque and Art competitions have highest mean age. In female group, Swimming and Gymnastics have lowest mean age while Alpinism and Art competitions have the highest.
- For "Height" attribute, in the male group, Basketball 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:

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

Age

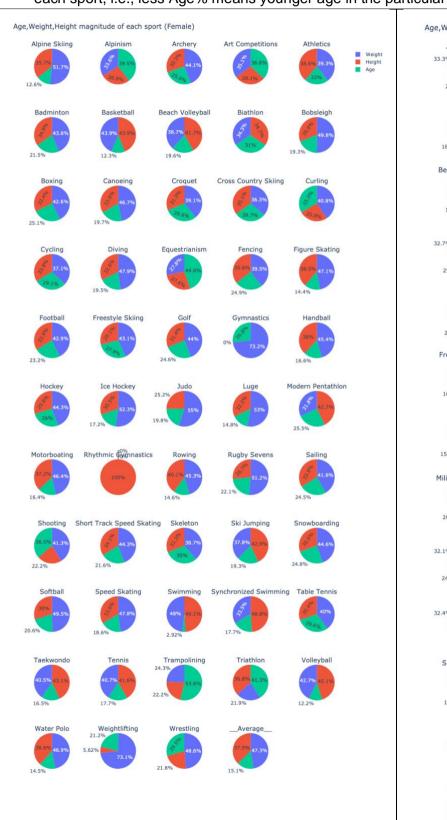
Age

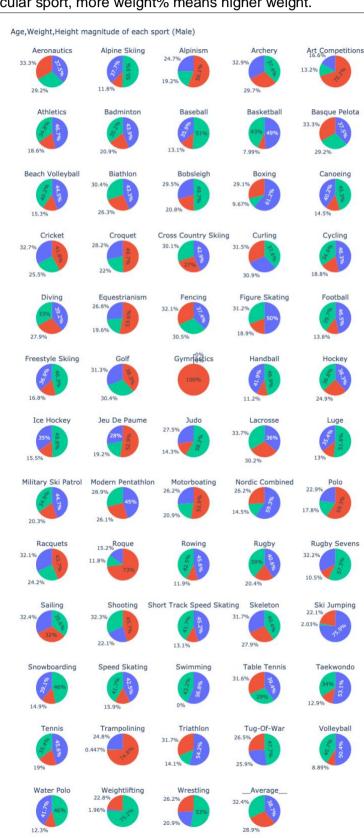


The graph provides the same information as the heatmap but in more details:

- 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.

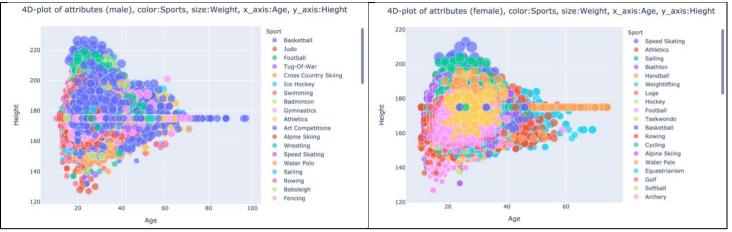
After that, I plot a multi pie chart, each sport is a pie, and attributes in the pie shows their contribution to each sport, i.e., less Age% means younger age in the particular sport, more weight% means higher weight.





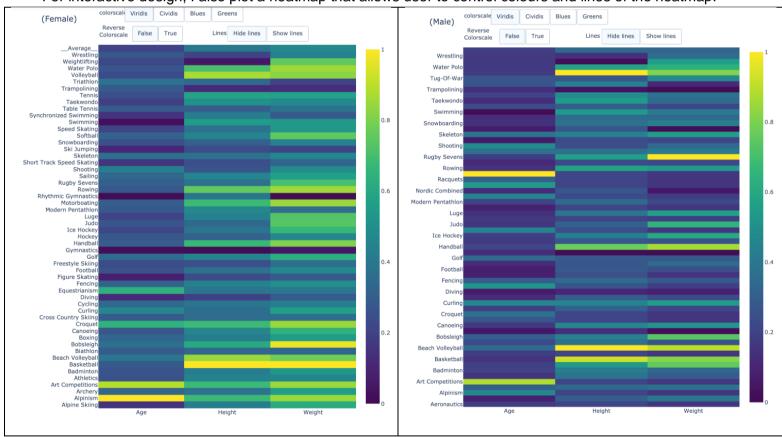
We can see that it is difficult to clearly compare each attribute between male and female, and it is hard to navigate between sports. However, it is easier to see how each attribute contribute to each sport.

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.

