Professor: *Dr. Mohammed Ayoub Alaoui Mhamdi*

Full Name: Alfred Ntiamoah

Number: 002317287

Signature: A.N

Bishop's University

CS 503 – Data visualization

Final exam

Fall 2021

December 20, 2021

Question under study

What type of sport is mostly influence by sex, age, height and weight?

My research is to understand the contributions of sex (gender), age, height and weight to the various sporting events. This may easily answer question like “Does weight influence the winning of tag-of-war sport?”.

This visualization will be done python libraries panda, matplotlib, plotly, numpy, seaborn and Scikitlearn

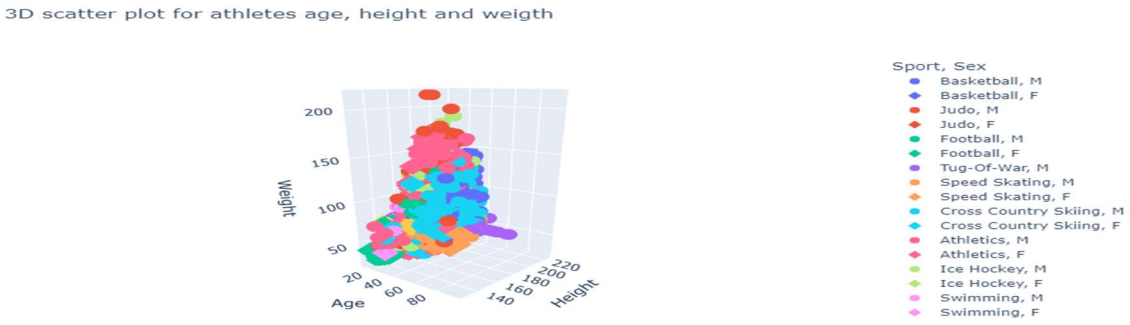
Data pre-processing

Most of the data give are clean except for age, height and weight. The age fields have about 9474 null values, the height field has 60171 null values and weight has about 62875 null values. These data are too much to ignore so I filled in these null values with the average of the provided data. Thus if age is not provided, I then average the provided ages and use that to replace the missing data.

Feature	**Average**
Age	25.56
Height	175.34
Weight	70.70

Visualizations

1. 3D scatter plot

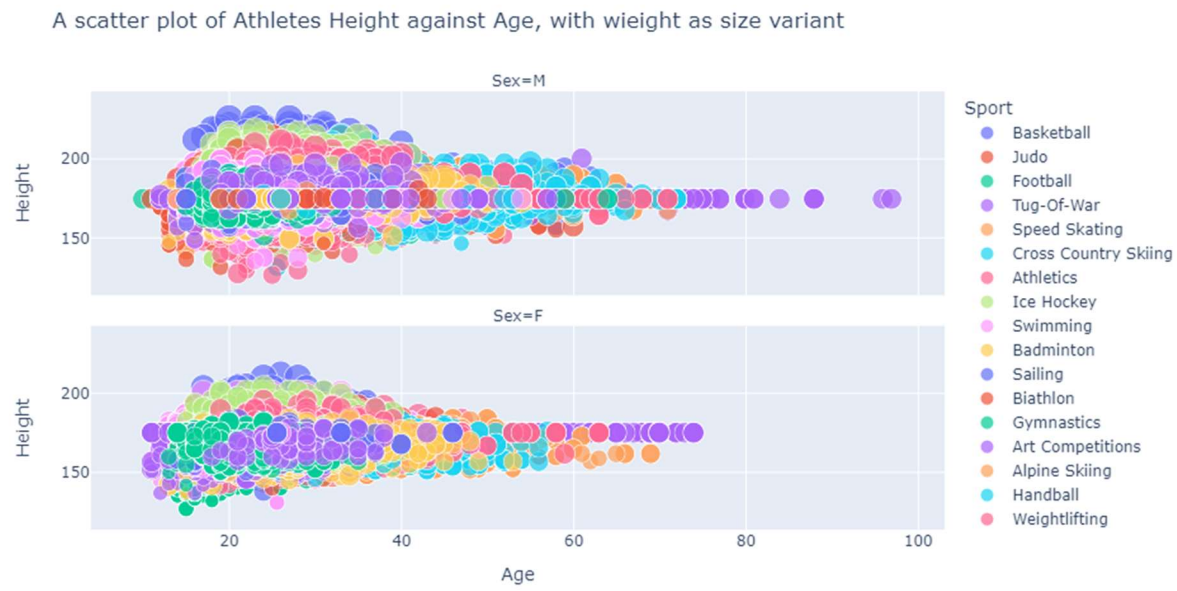


Using protly, 3d scatter plot is constructed for the data discussed above. Age, height, and weight are assigned to X, Y, Z axes respectively. Shape is used to differentiate between gender thus filled circle for male and a diamond for female. Color is used to distinguish the sports from each other.

From the graph, it can easily be seen that Judo and Gymnastic for both male and females are influence by weight among all the sports activities. Higher weights favored Judo while lower weight favored Gymnastic athletes. Athletes of higher age values participated in Art competitions while those of higher heights played Basketball.

The graph really showed the details nature of the data and answered the question posed above fully. This is best to represent the data among the three (3) graphs.

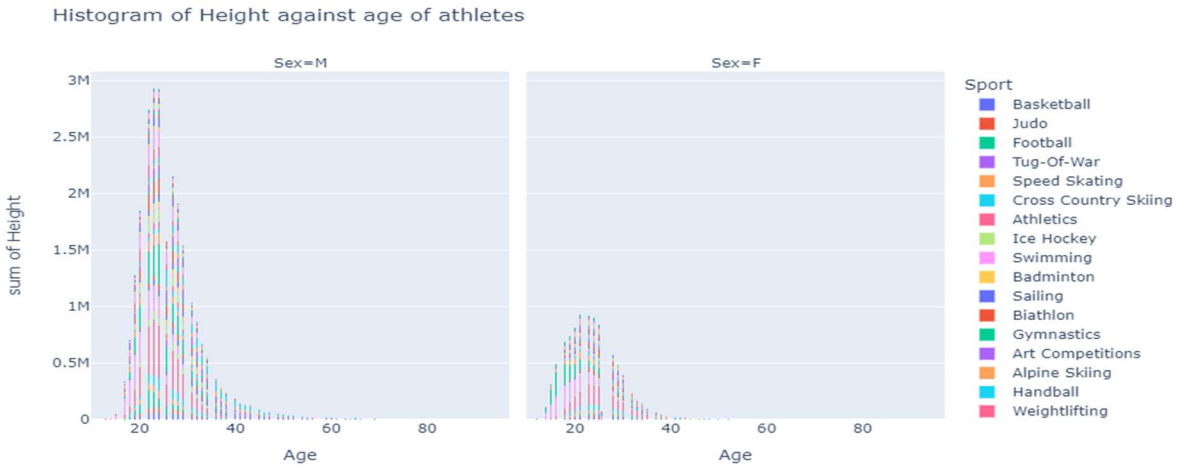
2. 2D Scatter plot with Facets



This graph show most details need to answer the question thus it was able to answer the question and also presented clear segregation between the sex. Though it performs well, there are several drawbacks that made me reject it as the best.

- The data was very clumsy at some portion of the graph. The plotly interactive library helped to resolve this issue by zooming. Since zooming only shows a small portion of the data, it makes it difficult to relate the data to other points that are out of view.
- The size parameter that was used to distinguish weight was near to uniform. All the data points appear nearly the same size, hence making it difficult to read.

3. Histogram (Stacked, Facet)



Also this, this graph was able to answer the question pose above. Though much cognitive effort is requiring to understand the relations between the data. It also offered a good segregation between the two gender.

It wasn't selected as the best because it requires more cognitive effort to explain.

Conclusion

The 3D scatter was the best among the three (3), followed by the Histogram and the 2d scatter coming last with the reasons already presented above.