SQL Capstone Project:
Analyzing the Effect of Age, Height, and Weight on Olympic Performance

Benjamin Prud'homme
October 23, 2022

Agenda

- Introduction
- Questions/Hypotheses
- Format of Data
- Data Cleaning / New Metrics
- Methodology/Technical Challenges
- Findings
- Summary
- Conclusion
- Citations

Introduction

- For this project, I am using the dataset entitled “Olympics Dataset - 120 years of data.” It includes records of athletes who have competed in the modern Olympics as of 2016, including:
  - The year, sport, and event they competed in
  - Physical properties like age/height/weight
  - What medal (if any) they won
- I seek to examine the effects that age, height, and weight have on performance in the Olympics, and how these effects differ between males and females as well as between various types of events.
- These findings will be particularly useful for personal trainers, fitness coaches, and health professionals partnered with SportsStats.

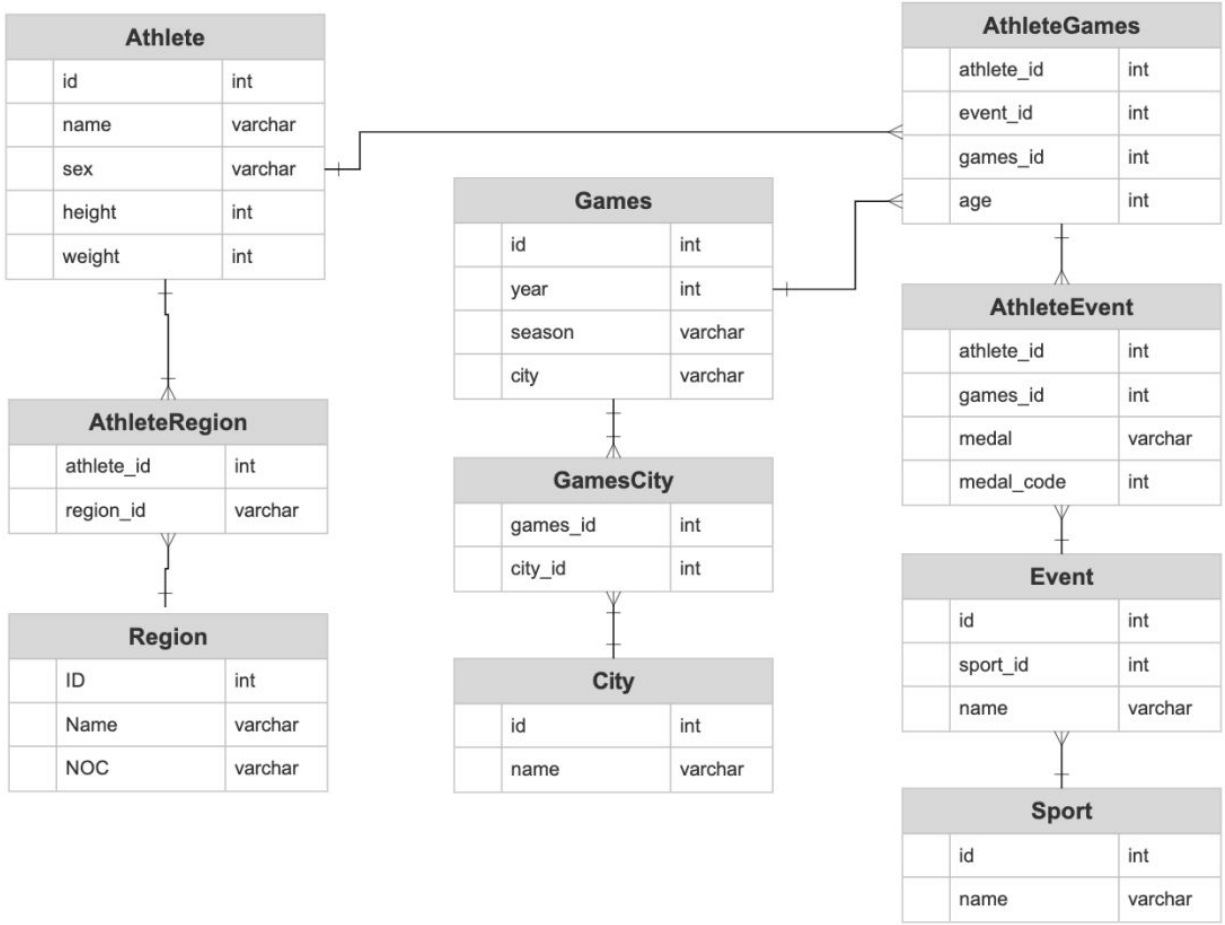
Questions/Hypotheses

- Initial questions:
  1. What effect does age have on performance?
  2. What effect do height, weight, and BMI (a way to express the ratio between them) have on performance?
  3. How do these effects differ for men and women, and between different events?
- Initial hypotheses:
  1. The best-performing athletes will be those in their mid-to-late 20's.
  2. BMI of the best athletes will generally be near the middle of the "healthy" range (about 20-23)
  3. Height and weight will be moderately to strongly positively correlated
  4. Height, weight, and BMI in women will be less than those in men
  5. In events requiring mostly strength (weightlifting, throwing, etc.), successful athletes will have higher height/weight/BMI values. In those requiring mostly speed/agility (track, gymnastics, etc.), they will have lower values.
- Later questions:
  1. How do the above effects differ between different categories of events?
  2. Which sports have the biggest differences between sexes/categories and/or deviate the most from the general trend, and why?

Format of Data

- Link to data: https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results
- The data came in 2 files:
  - athlete_events.csv
    - Categorical features: Name, Sex, Team, NOC (National Olympic Committee), Games, Season, City, Sport, Event, Medal
      - Medal = “Gold,” “Silver,” “Bronze,” or no value if the athlete did not win a medal
    - Numerical features: Age, Height (cm), Weight (kg), Year
      - Age values change with respect to year, but Height and Weight values are the same for all a particular athlete’s records
        - Description on Kaggle did not specify; assuming these are career averages
  - noc_regions.csv
    - Categorical features: NOC, region, notes
      - ‘NOC’ and ‘notes’ reflect territories as they were defined at the time an athlete competed at an Olympics. ‘region’ indicates what territories these areas belong to as of 2016.
- The relevant features for my analysis are: Sex, Age, Height, Weight, Medal
  - Plus a couple of new metrics (on the next slide)

Entity-Relationship Diagram



Data Cleaning / New Metrics

- Data Cleaning
  - Removed rows with missing age/height/weight values
- New Metrics
  - BMI = Weight (kg) / ((Height (cm))/100)^2
  - Medal Code: Gold = 1, Silver = 2, Bronze = 3, No medal = 4
  - Sport Category: Categorization of sports based on objectives, format, etc. (see below)
    - Separated Athletics into Track, Field, and All-Around, since they are very different

Category	Sports
Racing Sports Where the object is to get the fastest time	Alpine Skiing, Athletics (Track)*, Biathlon, Bobsleigh, Canoeing, Cross Country Skiing, Cycling, Luge, Motorboating, Rowing, Sailing, Speed Skating, Short Track Speed Skating, Swimming, Skeleton, Triathlon
Hi-Score Sports Where competitors take turns, and the best result wins	Athletics (Field)*, Archery, (Artistic) Gymnastics, Diving, Equestrianism, Figure Skating, Golf, Freestyle Skiing, Rhythmic Gymnastics, Shooting, Ski Jumping, Snowboarding, Swimming, Synchronized Swimming, Trampolining, Weightlifting
Versus Sports Where 2 athletes/teams compete directly against each other	Badminton, Baseball, Basketball, Beach Volleyball, Boxing, Curling, Fencing, Football, (Field) Hockey, Handball, Judo, Lacrosse, Ice Hockey, Rugby, Rugby Sevens, Softball, Table Tennis, Taekwondo, Tennis, Tug-Of-War, Volleyball, Water Polo, Wrestling
Combined Sports Combinations of sports from multiple of the above categories	Athletics (All-Around)*, Modern Pentathlon, Nordic Combined

Preview of Data

```
reduced_data.head(10)
```

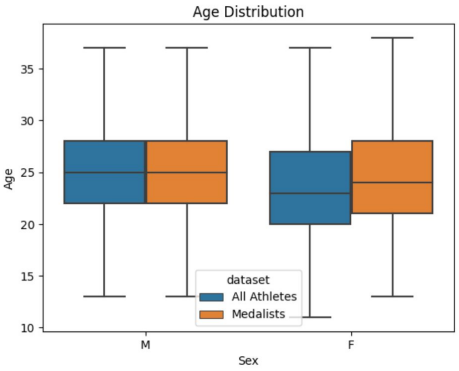
✓ 0.2s Python

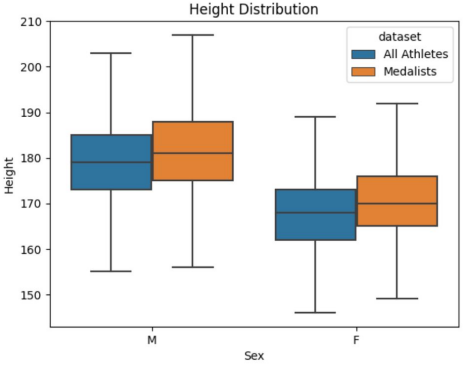
	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	region	notes	Medal Code	Sport Category	BMI
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN	China	NaN	4	Versus	24.691358
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN	China	NaN	4	Versus	20.761246
2	602	Abudoureheman	M	22.0	182.0	75.0	China	CHN	2000 Summer	2000	Summer	Sydney	Boxing	Boxing Men's Middleweight	NaN	China	NaN	4	Versus	22.642193
3	1463	Ai Linuer	M	25.0	160.0	62.0	China	CHN	2004 Summer	2004	Summer	Athina	Wrestling	Wrestling Men's Lightweight, Greco-Roman	NaN	China	NaN	4	Versus	24.218750
4	1464	Ai Yanhan	F	14.0	168.0	54.0	China	CHN	2016 Summer	2016	Summer	Rio de Janeiro	Swimming	Swimming Women's 200 metres Freestyle	NaN	China	NaN	4	Race	19.132653
5	1464	Ai Yanhan	F	14.0	168.0	54.0	China	CHN	2016 Summer	2016	Summer	Rio de Janeiro	Swimming	Swimming Women's 4 x 200 metres Freestyle Relay	NaN	China	NaN	4	Race	19.132653
6	3605	An Weijiang	M	22.0	178.0	72.0	China	CHN	2006 Winter	2006	Winter	Torino	Speed Skating	Speed Skating Men's 500 metres	NaN	China	NaN	4	Race	22.724403
7	3605	An Weijiang	M	22.0	178.0	72.0	China	CHN	2006 Winter	2006	Winter	Torino	Speed Skating	Speed Skating Men's 1,000 metres	NaN	China	NaN	4	Race	22.724403
8	3610	An Yulong	M	19.0	173.0	70.0	China	CHN	1998 Winter	1998	Winter	Nagano	Short Track Speed Skating	Short Track Speed Skating Men's 500 metres	Silver	China	NaN	2	Race	23.388687
9	3610	An Yulong	M	19.0	173.0	70.0	China	CHN	1998 Winter	1998	Winter	Nagano	Short Track Speed Skating	Short Track Speed Skating Men's 1,000 metres	NaN	China	NaN	4	Race	23.388687

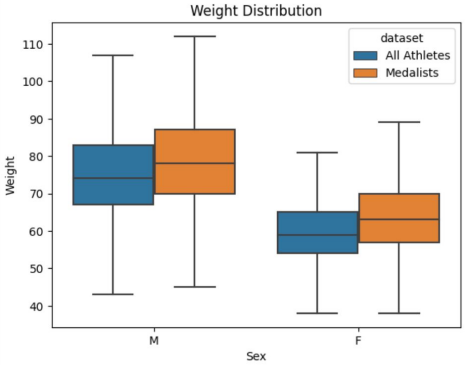
Methodology/Technical Challenges

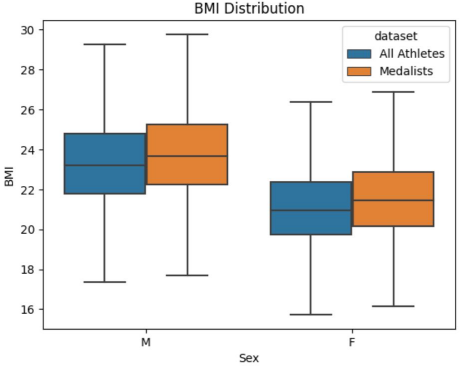
- The dataframes I used for deep analysis contain the following properties grouped by sport, including their category and...
  - Average metrics for all athletes and for medalists only, and the difference between these averages
  - Average metrics for men and for women, and the difference between these averages
  - Average height, actual average weight, predicted average weight (based on height with linear regression), and the difference between actual and predicted weight
- Challenges
  - This required the following intermediate steps:
    - Records containing only the records for medalists, or only the records for either men or women
    - Metrics grouped by sport for each of the two subsets of athletes I was looking to compare (all athletes vs medalists, or men vs women)
      - Final dataframes are the result of joining these on sport

Distributions: All Athletes vs. Medalists









Men

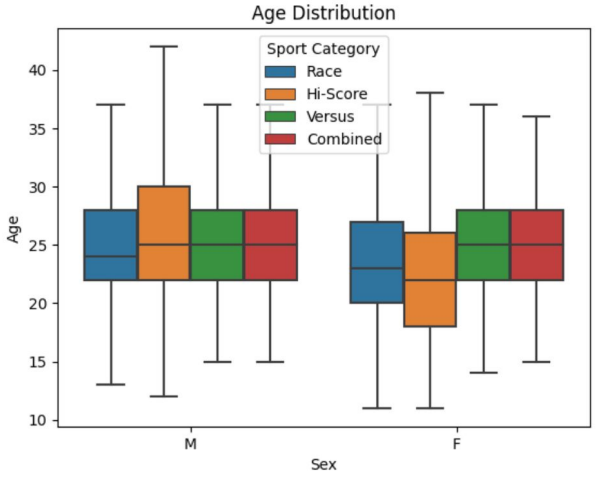
	dataset	AVG(Age)	AVG(Height)	AVG(Weight)	AVG(BMI)
0	All Athletes	25.667887	178.967548	75.795053	23.545946
1	Medalists	25.863130	181.353775	79.252080	23.964186

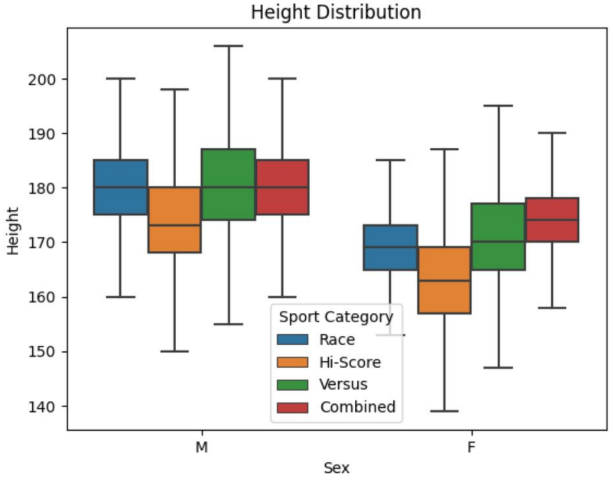
Women

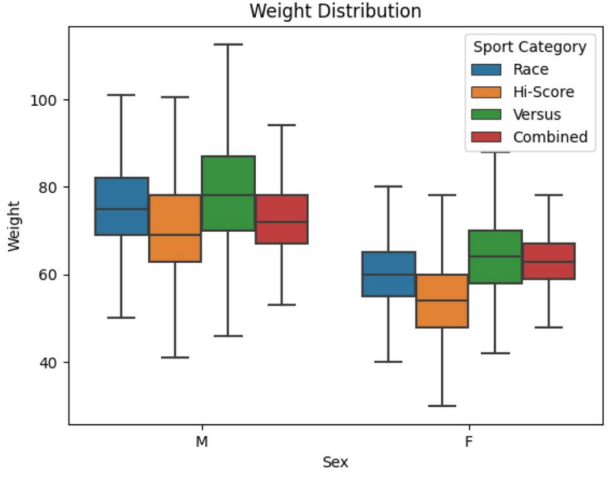
	dataset	AVG(Age)	AVG(Height)	AVG(Weight)	AVG(BMI)
0	All Athletes	23.781552	167.867261	60.027223	21.193275
1	Medalists	24.596152	170.538238	63.222663	21.631776

- Most male athletes are around 22-28 years old, 175-190 cm tall, weigh 67-87 kg, and have a BMI from 22-25.
- Most female athletes are around 20-27 years old, 160-172 cm tall, weigh 55-70 kg, and have a BMI from 20-23.
- Aside from the age of men (whose distribution is nearly equal), all distributions seem slightly higher for medalists vs the general population of athletes.
- This makes sense, since athletes who win medals are likely the ones who are more experienced, and have thus trained harder to get their bodies in shape.

Distributions: By Sport Category







Men

	Sport Category	AVG(Age)	AVG(Height)	AVG(Weight)
0	Combined	25.351005	177.862431	69.595978
1	Hi-Score	26.961565	174.303191	72.555794
2	Race	24.991929	180.591290	75.942359
3	Versus	25.612952	180.582244	78.774666

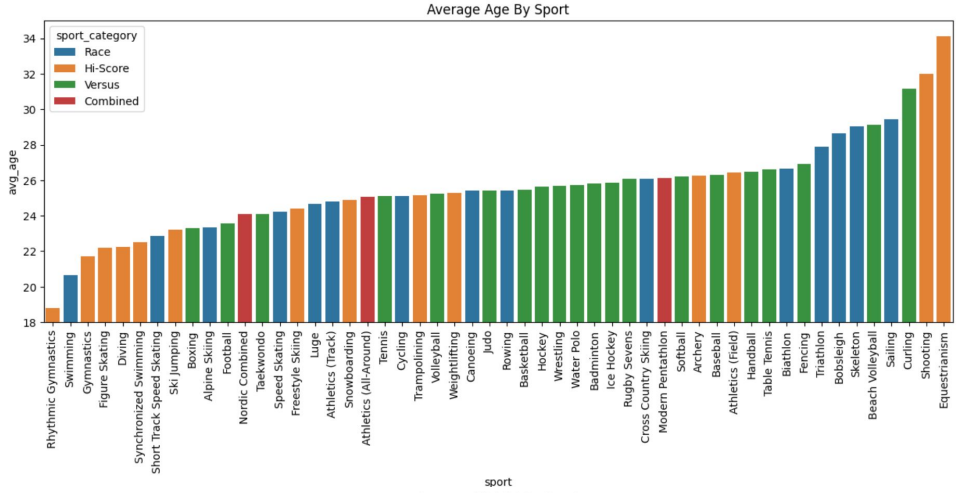
Women

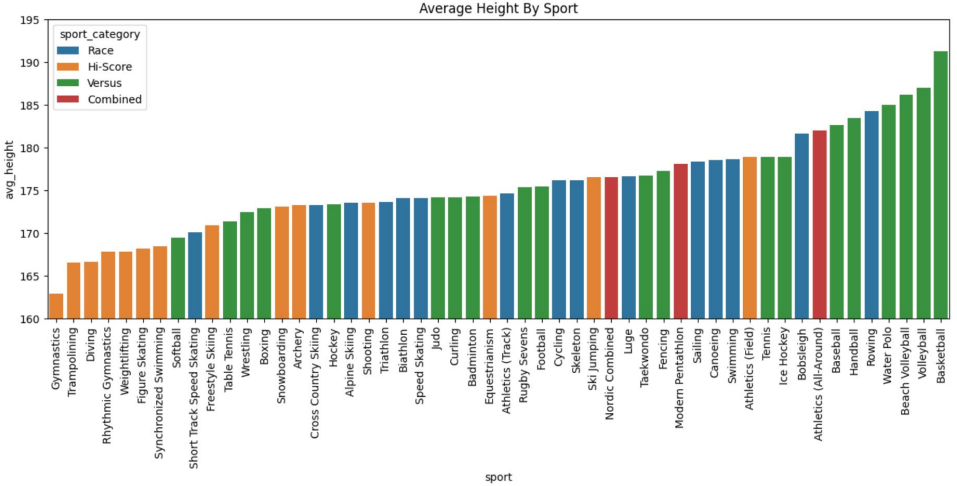
	Sport Category	AVG(Age)	AVG(Height)	AVG(Weight)
0	Combined	25.524390	170.073171	58.310976
1	Hi-Score	23.077226	162.981613	55.549977
2	Race	23.474879	169.223333	60.592578
3	Versus	25.393187	171.408947	64.821920

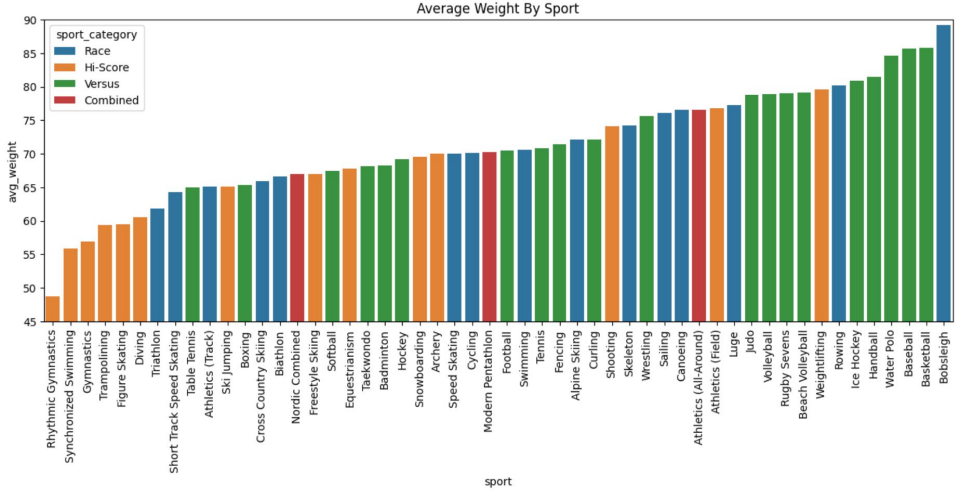
In general, the distributions of height and weight seem to be lowest for hi-score sports and highest for versus sports. This might be the case because:

- Certain hi-score events (Gymnastics, Figure Skating, Diving, Ski Jumping.) require flexibility, while others (Archery, Shooting, etc.) are more about strategy/accuracy than athletic ability
- In races, competitors are “active” for the whole race, requiring a constant supply of strength and/or endurance throughout.
- Many versus sports involve physically struggling with an opponent (fighting sports, Rugby, etc). Also, they often require quick reactions based on the actions of the opponent, which can be tougher to anticipate than say, the starting gun to a track race

Averages By Sport

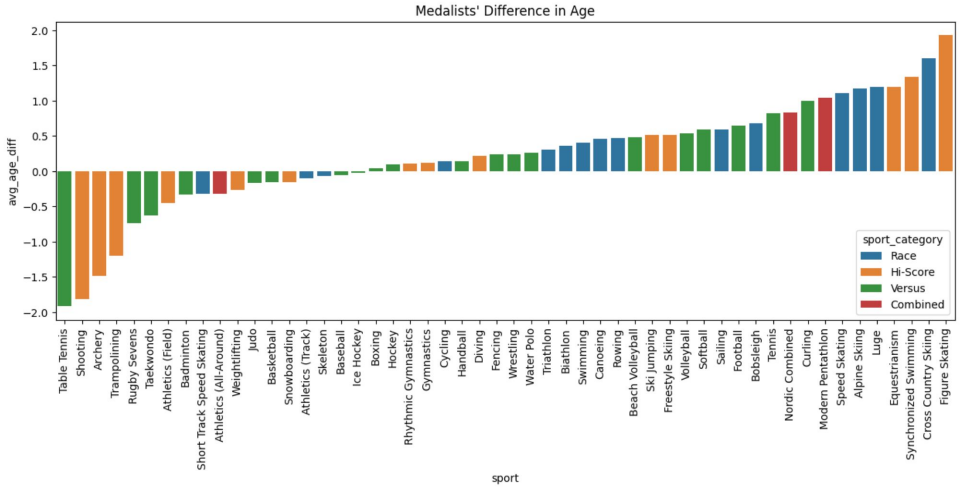


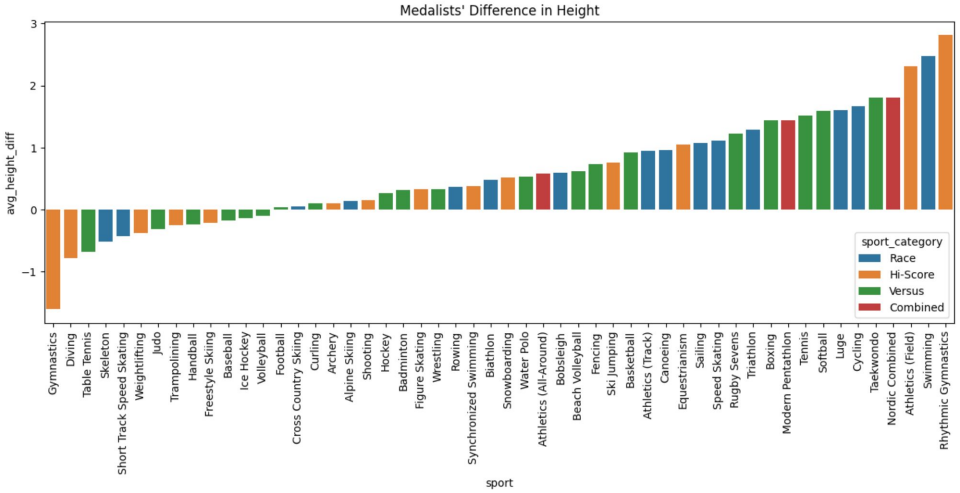


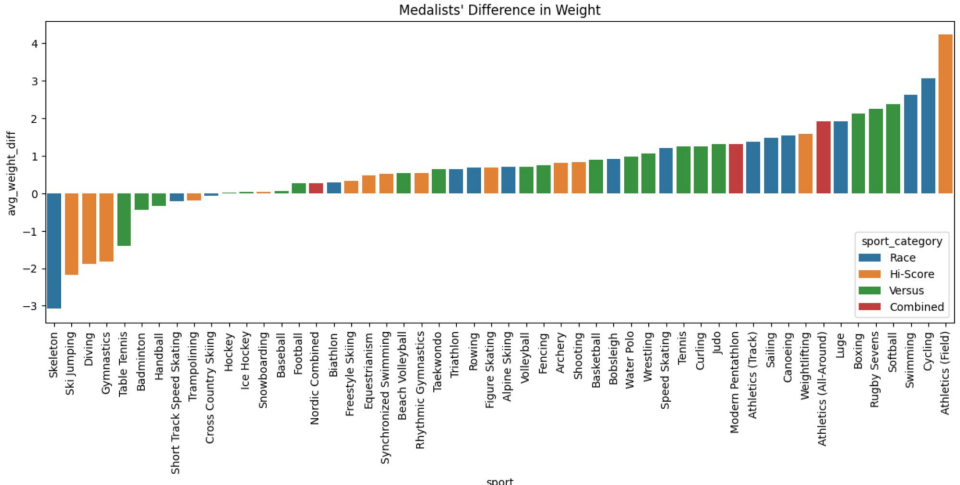


- Most of the sports with the tallest/heaviest athletes are team Versus sports, including: Basketball, Volleyball, Beach Volleyball, Water Polo, Handball
  - Understandable: height is a key advantage in these sports (shooting over an opponent in Basketball, leaping to block a spike in Volleyball, etc.), and height correlates to weight
- Sports with the shortest/lightest athletes include many Hi-Score sports demanding flexibility: Rhythmic Gymnastics, Sync. Swimming, (Artistic) Gymnastics, Trampolining, Figure Skating, Diving
  - Rhythmic and Artistic Gymnastics are also the sports with the youngest and 3rd-youngest competitors respectively.
    - Probably due to a large number of young, short female gymnasts
    - Rogers: These gymnasts are short because it allows them to generate more torque, light because it reduces the force required to overcome gravity, and young because of how long it takes to train effectively
- Athletes in Equestrian, Shooting, Curling, and Sailing are much older than the rest
  - Understandable: Curling is more strategic than athletic, and the others involve controlling an outside apparatus (horse/gun/boat), which is relatively non-strenuous
- A couple of surprises: Bobsleigh athletes are the heaviest, and Weightlifting athletes are among the lightest

Differences by Sport: Medalists vs. All Athletes

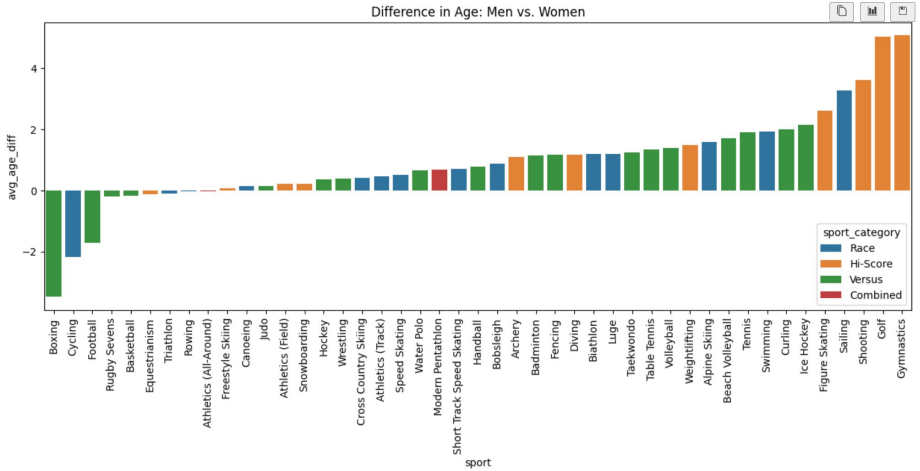


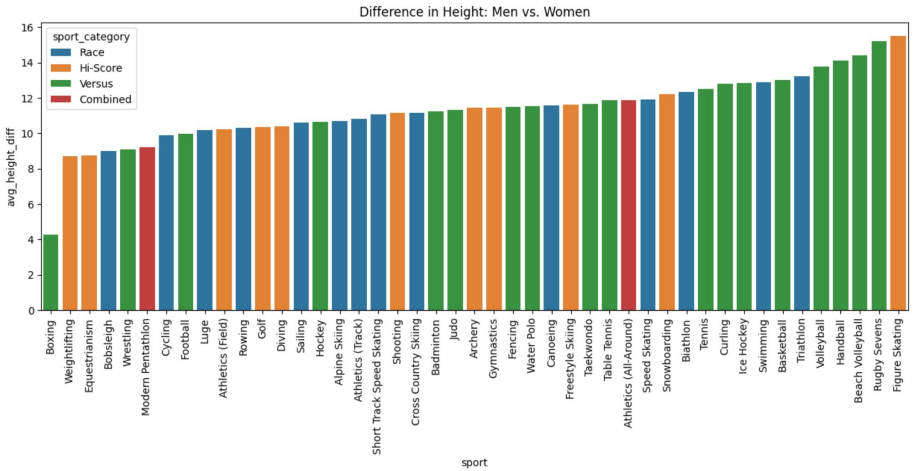


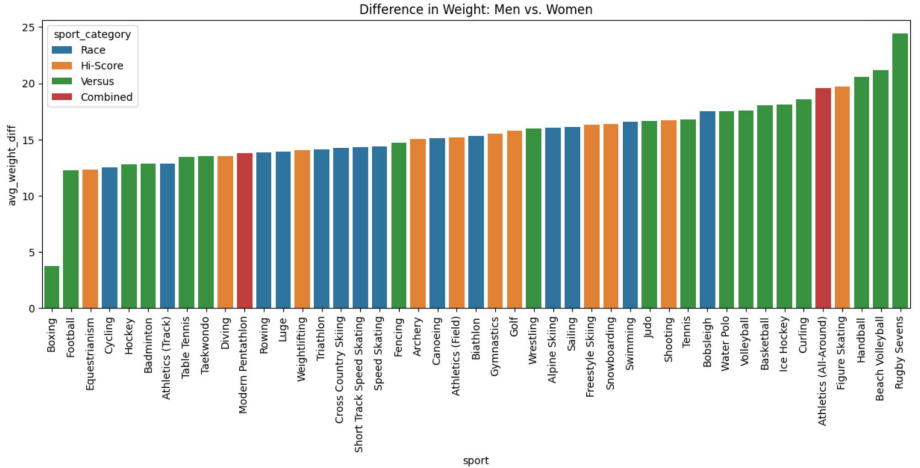


- Medalists are heaviest compared to average in Athletics (Field)
  - Understandable: includes throwing events (Javelin, Discus, Hammer, Shot Put) which are some of the purest strength Olympic events there are, and thus require a lot of muscle mass
- Medalists are 2nd- and 3rd-youngest compared to average in Shooting and Archery
  - Accuracy likely declines with age, plus t
- Medalists are shortest (and 4th-lightest) compared to average in (Artistic) Gymnastics (to be expected), but tallest compared to average in Rhythmic Gymnastics!
  - Velayos: Due to rhythmic gymnastics routines involving throwing and catching (balls, hoops, etc.), they are made easier by longer limbs
- Medalists are youngest, 3rd-shortest, and 5th-lightest compared to average in Table Tennis
  - Might appear surprising at first, but understandable: table tennis is extremely fast and requires agility and fast reflexes
- Medalists are 2nd-tallest and 3rd-heaviest compared to average in Swimming
  - A3 Performance: “Having a length advantage…gives them more surface area to propel themselves forward with

Differences by Sport: Men vs. Women

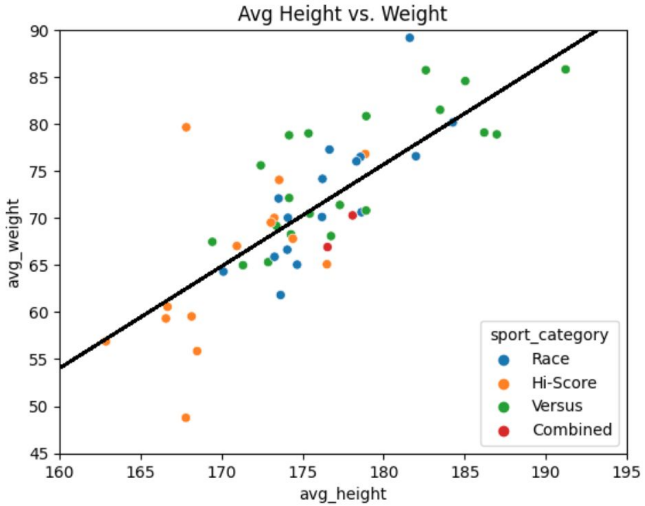


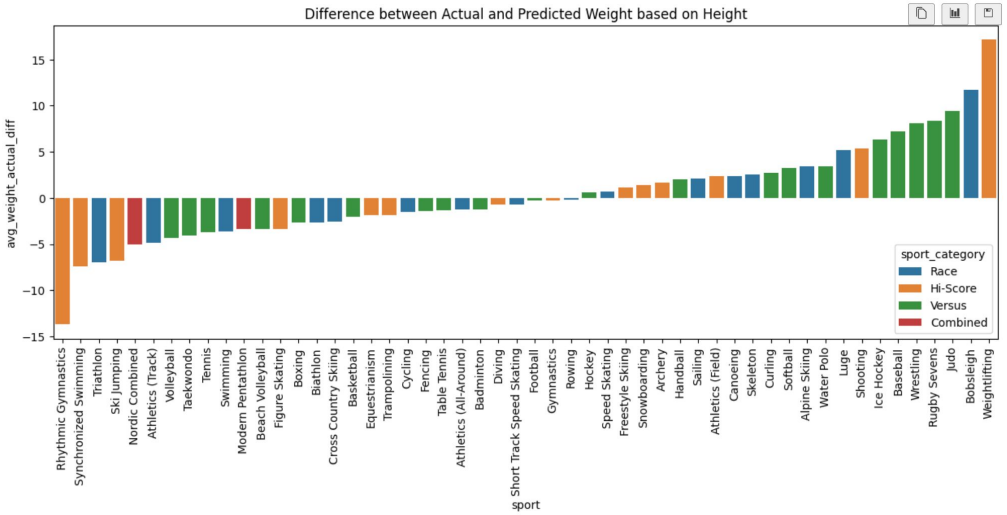




- Expectedly, men are consistently older, taller, and heavier than women within the same sports
- Outliers caused by non-physical factors:
  - Boxing: women are much less short/light compared to men than in other sports. Most likely because unlike other fighting sports (Judo, Taekwondo, Wrestling), the weight classes for men and women are the same (UXSquad)
  - Golf: similar difference. “The most concerning contributor…has to do with the massive wage gap”: 110 men vs only 11 women made more than $1 million in 2015 (Gallan)
  - Football: women are older than men. This is because men’s teams are restricted to under-23 athletes
- Gymnastics: men are ~5 years older than women
  - Rogers: “Much of men’s gymnastics involves a certain kind of upper-body strength…so it makes sense they tend to compete when their bodies are more fully formed and developed,” which doesn’t happen until well into the late teens and early 20’s
- Men tend to be the tallest/heaviest compared to women in Figure Skating and team sports (Rugby Sevens, Beach Volleyball, Handball, Volleyball, etc)
  - This is probably just due to biology

Height-Weight Correlation





- Moderate to strong positive correlation (r = 0.80)
- Outliers:
  - Weightlifting: approx. 16 kg heavier than expected
    - Reiss: the taller you are, the less you can lift as a percentage of your weight
  - Rhythmic Gymnastics: approx 13 kg lighter than expected
    - Velayos: rhythmic gymnasts have strict diets and put their bodies through extreme intensity, which causes height velocity to stop later in puberty for them - around the age of 18, rather than 15 for most women
  - Bobsleigh: approx. 11 kg heavier than expected
    - “A Short History of Weight Rules”: The more weight is in a bobsled, the faster it can go down the track. In fact, competitors with larger weights were consistently outperforming the rest of the competition until weight limits were introduced in the 1950s.
  - Judo, Rugby Sevens, Wrestling: approx. 8-9 kg heavier than expected
    - Understandable, since they involve physically struggling with opponents

Summary of Findings: Hypotheses

1. The best-performing athletes will be those in their mid-to-late 20's.
    a. Partially correct. The interquartile range of medalists (both male and female) included the early-to-mid 20's as well.
2. BMI of the best athletes will generally be near the middle of the "healthy" range (about 20-23)
    a. Correct for women, but incorrect for men, whose BMI interquartile range is closer to 22-25 (just on the border of being overweight, in fact)..
3. Height and weight will be moderately to strongly positively correlated
    a. Correct. My regression model yielded a coefficient of r=0.80.
4. Height, weight, and BMI in women will be less than those in men
    a. Correct.
5. In events requiring mostly strength (weightlifting, throwing, etc.), successful athletes will have higher height/weight/BMI values. In those requiring mostly speed/agility (track, gymnastics, etc.), they will have lower values.
    a. Mixed results. Sports like Gymnastics and Athletics (Field) met up with these expectations, but others such as Weightlifting (where athletes are short and heavy) and Athletics (Track) (which sticks close to the general age/height/weight trend) surprised me.

Summary of Findings: Later Questions

1. How do the above effects [of age/height/weight] differ between different categories of events?
    a. Generally, Hi-Score events have the youngest/shortest/lightest athletes. Versus events have the oldest/tallest/heaviest athletes, with Race events in the middle.
2. Which sports have the biggest differences between sexes/categories and/or deviate the most from the general trend, and why?
    a. I uncovered many insights; the most notable ones include:
        i. Men are oldest compared to women in Gymnastics and Golf. Women are oldest in Boxing, Curling, and Football.
        ii. In general, men are considerably taller and heavier than women in all sports, although not as much as usual in Boxing.
        iii. Team sport athletes are particularly tall/heavy
        iv. Athletes in sports demanding flexibility are short/light
        v. Gymnasts are particularly young/short/light, but it is the shortest artistic gymnasts and the tallest rhythmic gymnasts who are most successful
        vi. Weightlifters and bobsledders are particularly heavy considering their height
    b. Most of these differences are due to physical factors and/or the nature of particular sports, although a few are caused by restriction or classification of competitors, or even social/cultural issues (in the case of Golf).

Conclusion

- Recommendations
  - Personal trainers and Olympic coaches should adapt their training regimes in a way that is conducive to success in one's sport of interest
  - They could also use this data, along with the facts that support it, to more effectively advise those who have not yet committed to a particular sport, considering their current physical state
- Next steps for analysis could include:
  - Analyzing trends over time
    - Have athletes gotten younger/older, shorter/taller, lighter/heavier - in general and in given sports?
  - Investigating country trends
    - Which nations are most successful - in general and in given sports?
    - What factors (government, economy, etc.) could be responsible for these differences?

Citations

A3 Performance. “Here Is Why Swimmers Are so Tall, and What to Do If You Are Not.” A3 Performance, https://www.a3performance.com/blogs/a3-performance/swimmers-tall-and-short.

Gallan, Daniel. “Why Young Women Dominate Golf but Struggle with Longevity - Conqa Group: Lead • Grow • Connect.” CONQA Group, CONQA Group, 8 Nov. 2017, https://www.conqagroup.com/blog/womens-golf-lpga-pga-ashleigh-simon-longevity-young-athletes-dominate.

Reiss, Sam. “Why Are Weightlifters so Short? the Height-Lifting Connection, Explained.” Inverse, Inverse, 29 May 2021, https://www.inverse.com/mind-body/leg-day-observer-weightlifting-height.

Rogers, Joshua. “Why Are Gymnasts so Young? How Team USA and the Tokyo Olympics Is Challenging the Status Quo.” The Focus, 29 July 2021, https://www.thefocus.news/sports/olympics/why-are-gymnasts-so-young/.

“A Short History of Weight Rules.” Bobskeleton.org.uk, 1 Nov. 2018, https://bobskeleton.org.uk/?p=73.

UXSquad. “The Difference in Rules of Mens and Womens Boxing.” Unorthodoxx, Unorthodoxx, 2 May 2021, https://www.unorthodoxx.co.uk/post/the-difference-in-rules-of-mens-and-womens-boxing.

Velayos, Diana. “Why Are Rhythmic Gymnasts so Tall? How Tall Are Gymnasts?” Diario AS, 6 Aug. 2021, https://en.as.com/en/2021/08/06/olympic_games/1628240586_789360.html.