

SQL for Data Science Capstone - Milestone 2

Francesco Aldo Tucci, 2021-12-27

My analysis, as briefly summarised in the Week 1 Milestone Questions & Hypotheses sections, is mainly concerned about the age and fitness profile of Olympic winners.

The Olympic *athlete_events* data set contains information about athletes who took part in the Olympic games. The data starts from the very first edition of the modern Olympic games (Athens 1896) and covers up to the 2016 edition (included), though for our purposes the actual data will be a slightly selected subset, since complete information about age, height and weight is less frequent for older events. I acknowledge this limitation, though addressing it goes well beyond the scope of this analysis. Information about the Olympic games themselves (year, NOC code, city, etc.) is complete. For each edition, the data set contains the name, sex, age, height, and weight of the athletes participating in each event (identified by the *Sport* name as well as the complete name of the event), together with the “banner” of the nation he/she was competing under (variable *Team*), and which medal he/she won, if any. BMI (Body Mass Index), defined as the ratio between weight in kilograms and square of height in centimeters, times 10,000, is computed for each athlete:

$$BMI = \frac{Weight(kg)}{Height(cm)^2} * 10000$$

I start by reporting a simple summary descriptive statistics of the quantitative variables that will be the main focus of my work. Since each sport has its own characteristics, each athlete-edition-sport combination was treated as a different observation (which means in practice that some medal winners will have the same age and BMI just because it is the same athlete).

Here I look at the usual measures of central tendency (mean, median and mode), plus descriptive statistics which give an idea about the variability in the variables (such as min, max, range and inter-quartile range, and standard deviation). I also report the number of missing observations, so as to get a first grasp on how sparse the data set is.

Table 1 - Descriptive Statistics (All Observations)

All	Age	Height (cm)	Weight (kg)	BMI
Min.	10	127.0	25.0	8.36
1st Q.	21	168.0	60.0	20.96
Median	24	175.0	70.0	22.53
Mean	25.56	175.3	70.7	22.78
3rd Q.	28	183.0	79.0	24.21
Max.	97	226.0	214.0	63.90
Range	87	99.0	189.0	55.54
IQR	7	15.0	19.0	3.25
Std. Dev.	6.3936	10.5185	14.3480	2.9121
N° Obs.	271116	271116	271116	271116
N° Missing	9474	60171	62875	64263

Next, I present descriptive statistics of the main variables of interest with some preliminary breakdown by *Season* (Summer vs Winter Olympics). Then, I will focus on Olympic medals winners, distinguishing between male and female athletes, since there are sex differences in BMI as well as age profile¹.

¹Some events, such as female gymnastics, tend to have younger winners compared to male gymnastics, due to characteristics intrinsic to the sport itself.

To give a first idea on why sex differences are likely to matter, I first present aggregate descriptive statistics (all athletes vs winners only), grouped by *Season*, then compute them again by sex group².

Table 2 - Descriptive Statistics (Summer Games, All Observations)

Summer all	Age	Height (cm)	Weight (kg)	BMI
Min.	10	127.0	25.0	8.36
1st Q.	21	168.0	60.0	20.96
Median	24	175.0	70.0	22.53
Mean	25.67	175.5	70.69	22.78
3rd Q.	28	183.0	79.0	24.21
Max.	97	226.0	214.0	63.90
Range	87	99.0	189.0	55.54
IQR	7	15.0	19.0	3.25
Std. Dev.	6.6998	10.9147	14.8039	2.9982
N° Obs.	222552	222552	222552	222552
N° Missing	9189	51857	53854	55191

Table 3 - Descriptive Statistics (Summer Games, Winners Only)

Summer win	Age	Height (cm)	Weight (kg)	BMI
Min.	10	136.0	28.0	13.49
1st Q.	22	170.0	63.0	21.28
Median	25	178.0	73.0	22.86
Mean	25.9	177.9	73.95	23.11
3rd Q.	29	185.0	83.0	24.49
Max.	73	223.0	182.0	56.49
Range	63	87.0	154.0	43
IQR	7	15.0	20.0	3.21
Std. Dev.	6.075	11.2161	15.3568	3.0165
N° Obs.	34088	34088	34088	34088
N° Missing	721	7990	8490	8732

²The above represents an additional dimension of the analysis, not considered in the Proposal, which emerged naturally after reflecting on the links between sex, age, BMI, and sport characteristics.

Table 4 - Descriptive Statistics (Winter Games, All Observations)

Winter all	Age	Height (cm)	Weight (kg)	BMI
Min.	11	137.0	32.0	13.77
1st Q.	22	168.0	62.0	21.30
Median	24	175.0	70.0	22.84
Mean	25.04	174.6	70.76	23.05
3rd Q.	28	181.0	79.0	24.54
Max.	58	211.0	145.0	39.54
Range	47	74.0	113.0	25.77
IQR	6	13.0	17.0	3.24
Std. Dev.	4.7777	8.5982	12.2133	2.4971
N° Obs.	48564	48564	48564	48564
N° Missing	285	8314	9021	9072

Table 5 - Descriptive Statistics (Winter Games, Winners Only)

Winter win	Age	Height (cm)	Weight (kg)	BMI
Min.	13	145.0	36.0	16.66
1st Q.	23	169.0	63.0	21.51
Median	26	176.0	72.0	23.23
Mean	26.06	175.5	72.81	23.43
3rd Q.	29	182.0	82.0	25.21
Max.	58	201.0	130.0	34.07
Range	45	56.0	94.0	17.41
IQR	6	13.0	19.0	3.70
Std. Dev.	4.8611	8.7438	13.0338	2.6261
N° Obs.	5695	5695	5695	5695
N° Missing	11	721	837	855

Next, as anticipated, I present the descriptive statistics by *Sex*. For convenience purposes, I focus on winners only, at this point.

Table 6 - Descriptive Statistics (Summer Winners, Females)

Summer win F	Age	Height (cm)	Weight (kg)	BMI
Min.	11	136.0	28.0	13.49
1st Q.	21	165.0	57.0	20.08
Median	24	171.0	63.0	21.43
Mean	24.38	171.1	63.58	21.59
3rd Q.	27	177.0	70.0	22.79
Max.	69	210.0	167.0	50.98
Range	58	74.0	139.0	37.49
IQR	6	12.0	13.0	2.71
Std. Dev.	5.3452	9.3079	11.0913	2.5863
N° Obs.	9442	9442	9442	9442
N° Missing	17	685	775	777

Table 7 - Descriptive Statistics (Summer Winners, Males)

Summer win M	Age	Height (cm)	Weight (kg)	BMI
Min.	10	140.0	37.0	15.85
1st Q.	22	175.0	70.0	22.20
Median	25	182.0	78.0	23.62
Mean	26.50	181.4	79.27	23.91
3rd Q.	29	188.0	88.0	25.01
Max.	73	223.0	182.0	56.49
Range	63	83.0	145.0	40.64
IQR	7	13.0	18.0	2.81
Std. Dev.	6.2382	10.4930	14.5004	2.9181
N° Obs.	24646	24646	24646	24646
N° Missing	704	7305	7715	7955

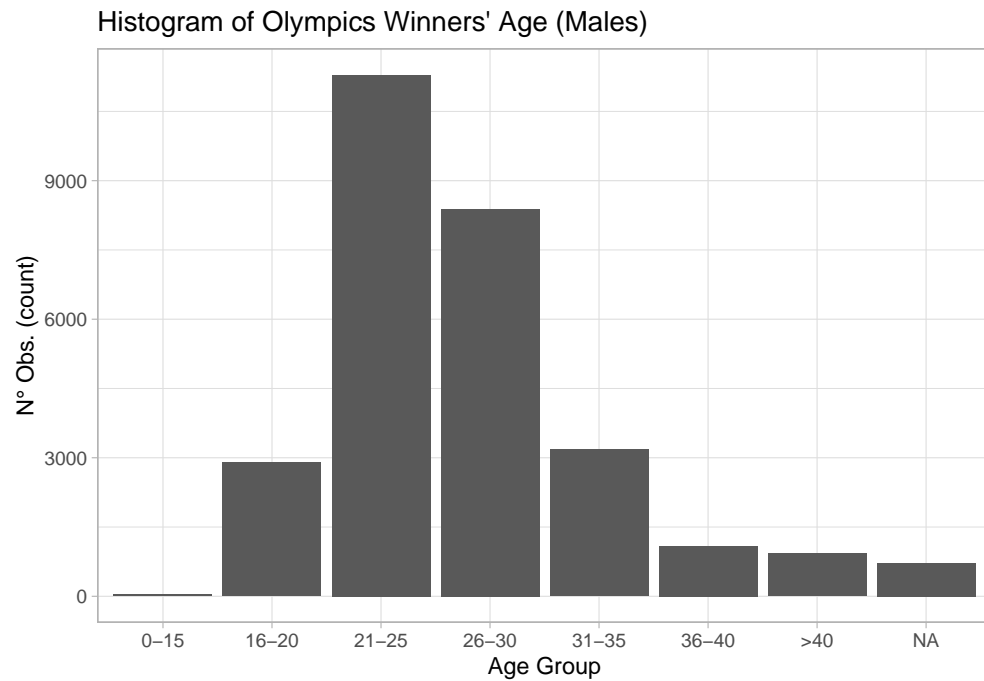
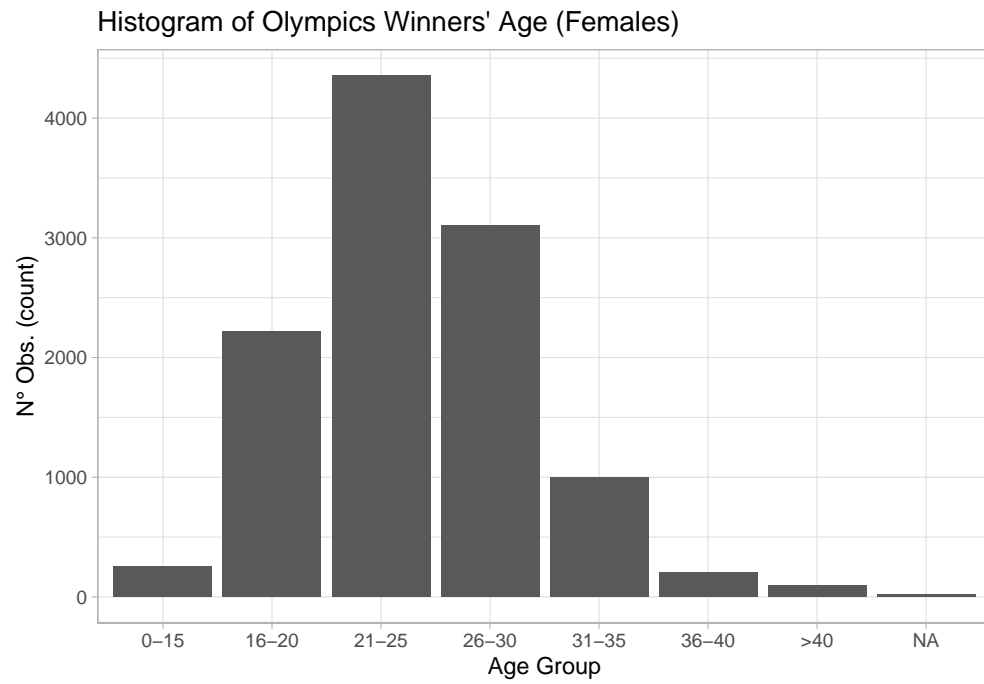
Table 8 - Descriptive Statistics (Winter Winners, Females)

Winter win F	Age	Height (cm)	Weight (kg)	BMI
Min.	13	145.0	36.0	16.66
1st Q.	22	163.0	56.0	20.45
Median	25	167.0	61.0	21.76
Mean	25.19	167.3	61.38	21.85
3rd Q.	28	172.0	66.0	23.05
Max.	46	187.0	95.0	32.87
Range	33	42.0	59.0	16.21
IQR	6	9.0	10.0	2.60
Std. Dev.	4.8249	6.2086	8.0023	2.0625
N° Obs.	1811	1811	1811	1811
N° Missing	0	90	126	126

Table 9 - Descriptive Statistics (Winter Winners, Males)

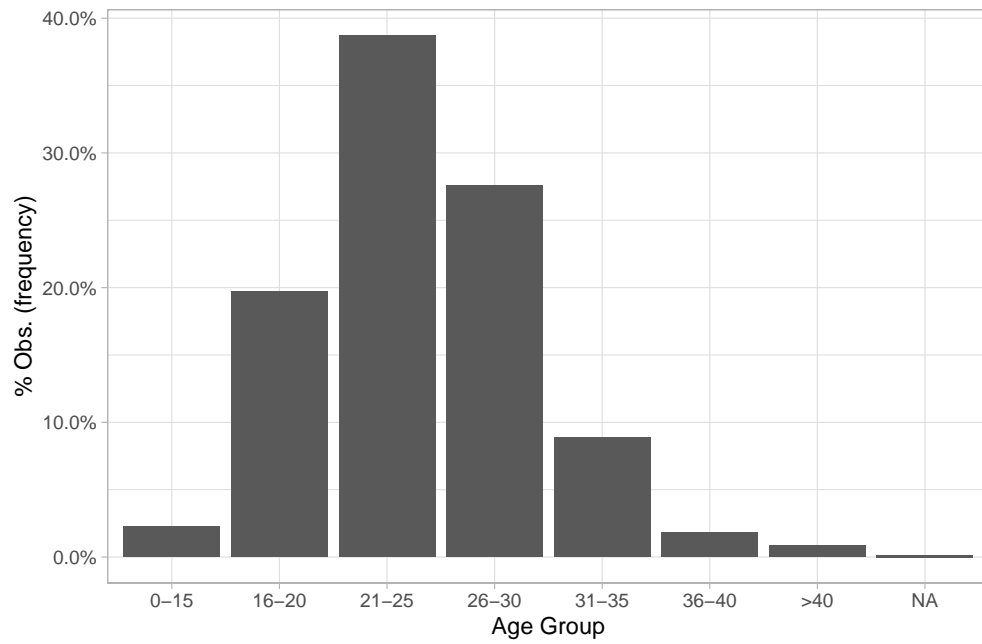
Winter win M	Age	Height (cm)	Weight (kg)	BMI
Min.	14	152.0	50.0	17.24
1st Q.	23	176.0	71.0	22.54
Median	26	180.0	79.0	24.30
Mean	26.47	179.9	78.88	24.26
3rd Q.	29	184.0	86.0	25.93
Max.	58	201.0	130.0	34.07
Range	44	49.0	80.0	16.83
IQR	6	8.0	16.0	3.39
Std. Dev.	4.8241	6.4138	10.9489	2.5074
N° Obs.	3884	3884	3884	3884
N° Missing	11	631	711	729

Next, I present some simple histograms for *Age* and *BMI*, the main variables of interest for my analysis.



To help comparison between male and female winners, given the different sample sizes, I also report frequency histograms for the age groups.

Histogram of Olympics Winners' Age (Females)



Histogram of Olympics Winners' Age (Males)

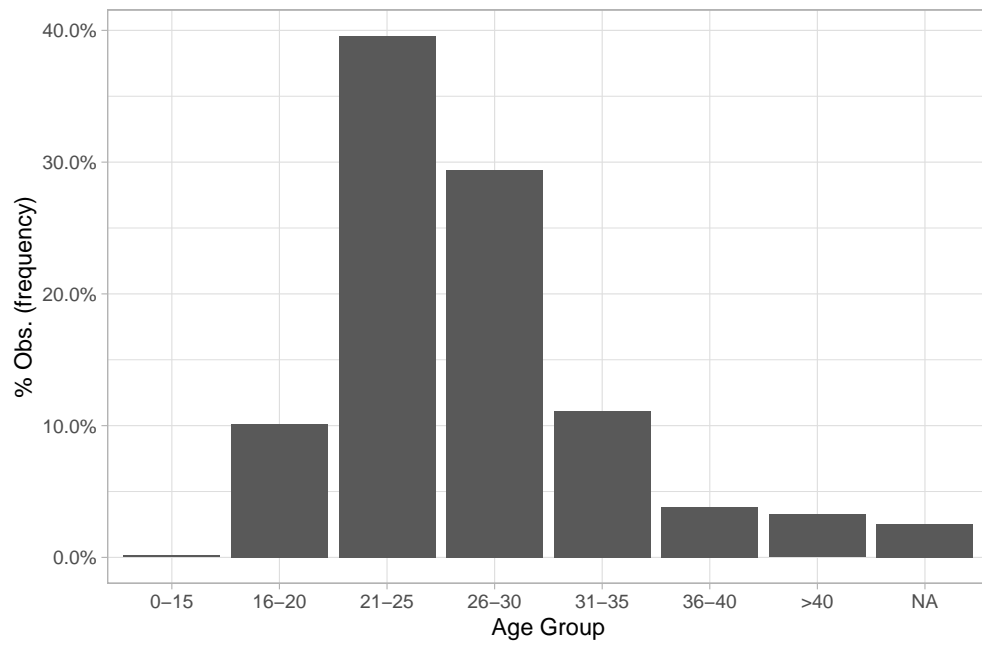
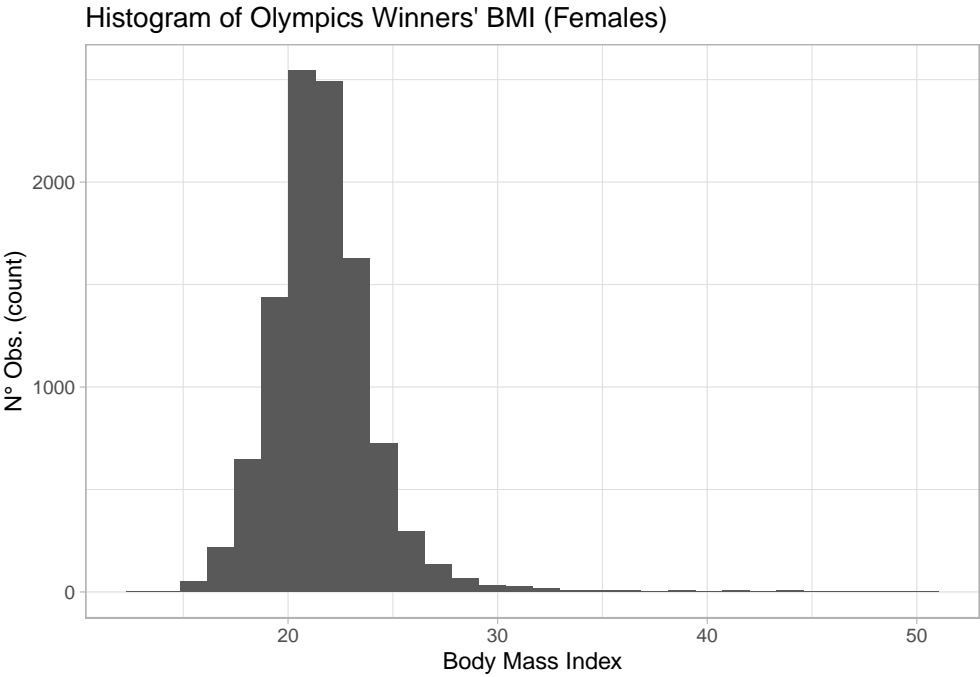
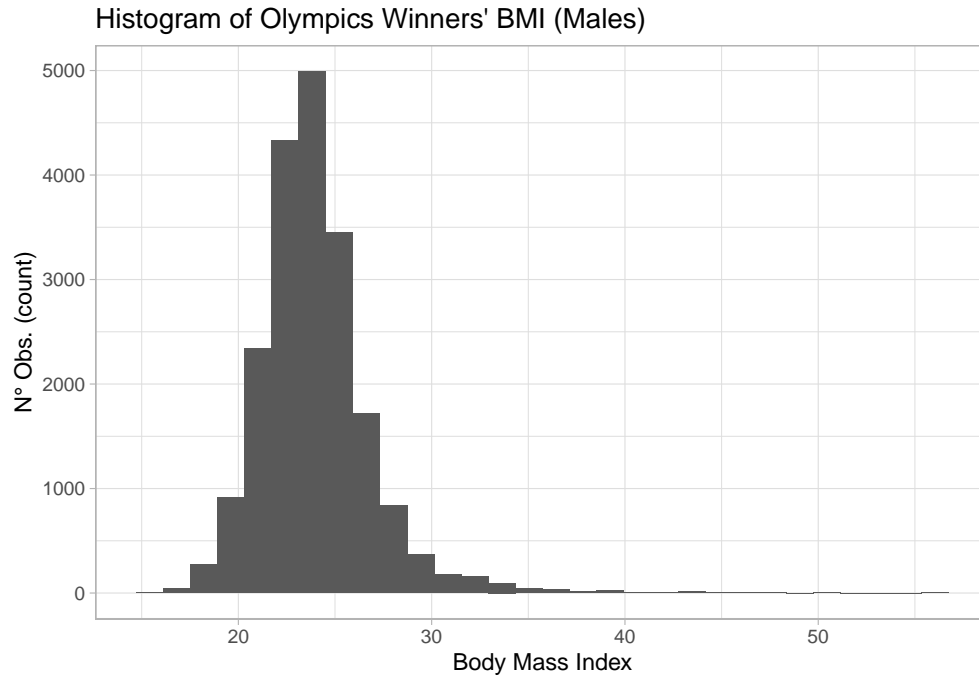


Table 10 - Age Group Composition of Winners, Percentages (M/F)

	Males	Females
0-15	0.14%	2.26%
16-20	10.15%	19.75%
21-25	39.59%	38.72%
26-30	29.40%	27.56%
31-35	11.14%	8.89%
36-40	3.82%	1.83%
> 40	3.26%	0.85%
NAs	2.51%	0.15%





Here, I present a correlation matrix. Each entry represents the Pearson correlation coefficient, computed using pairwise complete observations³, rounded to two decimal points, together with a symbol for statistical significance⁴.

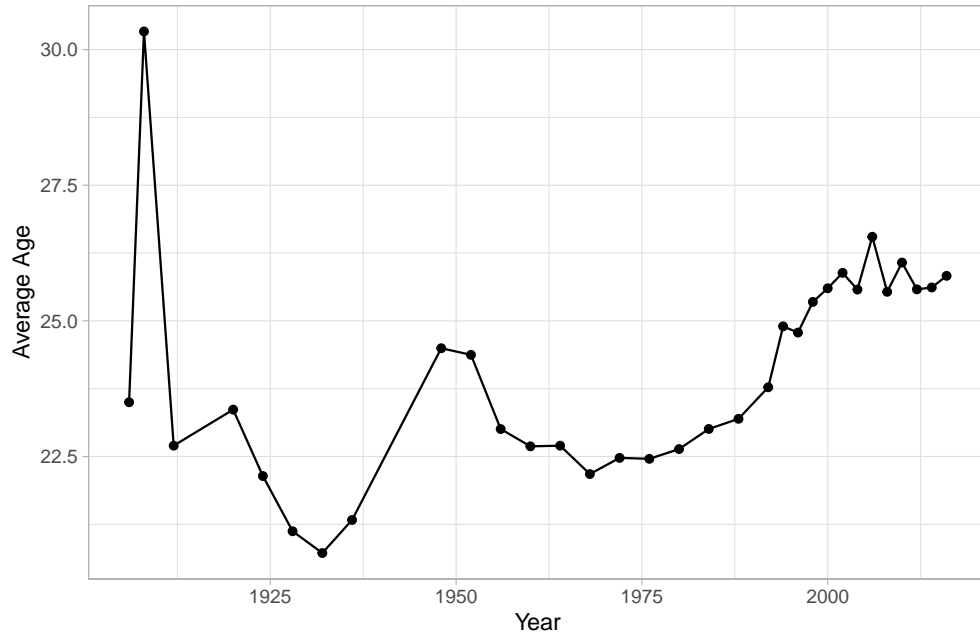
Table 11 - Correlation Matrix with Statistical Significance

	BMI	Height	Weight	Age	Male Dummy	Summer Dummy
BMI	1.00					
Height	0.33***	1.00				
Weight	0.82***	0.80***	1.00			
Age	0.18***	0.09***	0.16***	1.00		
Male Dummy	0.37***	0.46***	0.50***	0.15***	1.00	
Summer Dummy	-0.04***	0.08***	0.03***	-0.01^	0.003***	1.00

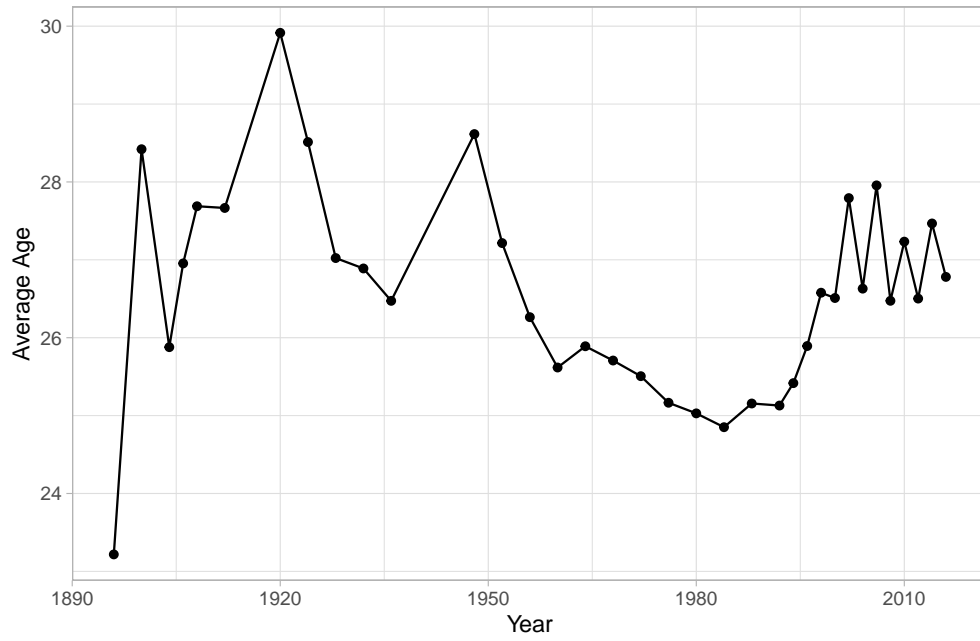
³While not reported, all correlation pairs use a sample size of at least 30,000 observations out of the 39,783 total.

⁴The following convention has been adopted: ^ 10%, * 5%, ** 1%, and *** 0.1% or less.

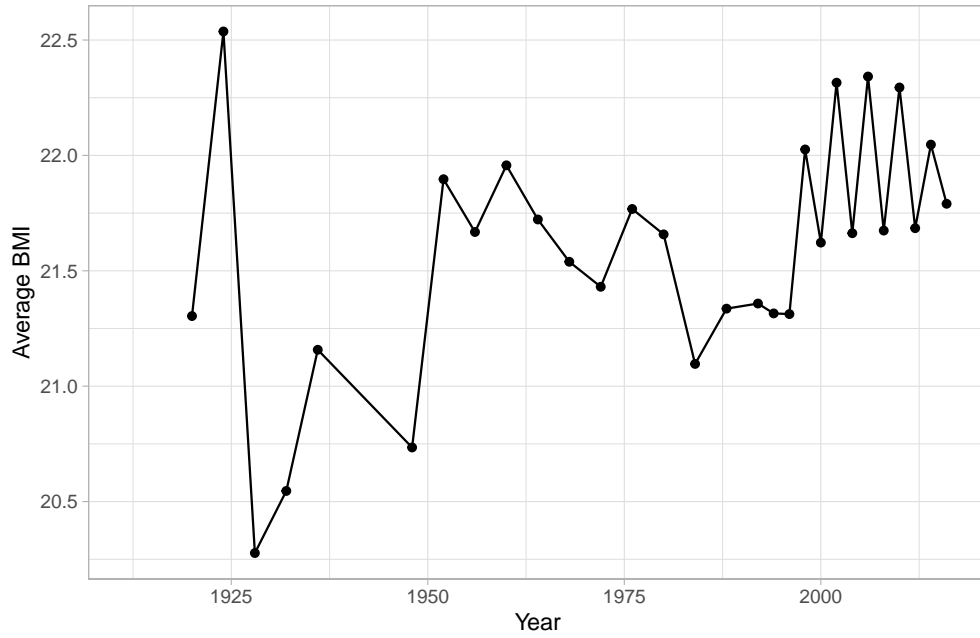
Winners' Age Over Time (Females)



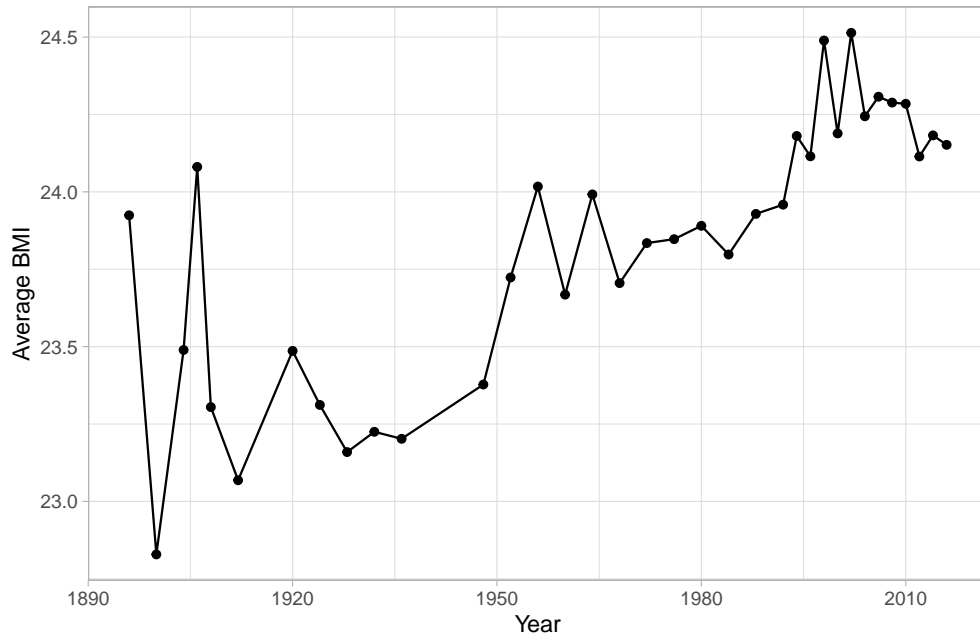
Winners' Age Over Time (Males)



Winners' BMI Over Time (Females)



Winners' BMI Over Time (Males)



Summary of Findings

The descriptive statistics show that differentiating between *Sex* and *Season* (i.e., implicitly, by type of sports) is sensible.

Differences in average values, as well as in variability, suggest the presence of systematic differences in the age profile and BMI values of the Olympic Games winners.

Overall, females tend to win while younger, likely due to the characteristics of some particular sports, e.g. gymnastics. While only 14 in 10,000 (0.14%) male winners are less than 16 years old, in relative terms there are more than 16 times as much very young female winners (2.26%). On the other hand, more than 7% of male winners are over 35, while only 2.7% of women are. As expected, most athletes who got a medal are in their primes: the bulk of winners is in the 21-30 category (about 70% of men and 66.3% of women).

As known from the literature, men have on average higher BMIs compared to women. Variability is somewhat higher as well. These stylized facts are reflected in the statistics for Olympic winners.

These first results allow to provide preliminary answers to (some) of the questions I raised in the Milestone 1 Report.

As far as 1.1⁵ is concerned, we see that female winners' average age has increased almost monotonically in the last four decades, inverting the descending trend that was going on from the '50s up until 1975. A somewhat similar (even more pronounced) U-shaped pattern applies to male athletes: average winner age had been trending down for about 3 decades (1950-1980) before it started increasing again. The last 10-20 years saw some oscillations, mostly due to the differences between "Summer" and "Winter" Olympics. On this matter, the next step will be to formally account for this apparent "seasonality" component of the time series for average age, which can be argued is actually due to Summer and Winter sports having different characteristics as well as physical requirements. The hypothesis of increasing winners' age seems to be verified. As written in the previous report, the mechanism identified is that modern sport medicine (including diet, rehabilitation, etc.) allows athletes to stay at the top of their games for longer periods of time. While these data do not allow to formally test for such a mechanism, it is a sensible enough rationalization of the observed trends.

To answer 1.2⁶, and actually go a little bit beyond that, a correlation matrix is shown. The statistical significance of the computed correlations has been formally tested⁷. All the correlations have the expected sign and are statistically significant (though the one between the seasonal dummy, equal to one for the Summer games, and *Age* has a p-value of 0.0561). In particular, there is a statistically significant, positive linear correlation between *Age* and *BMI*. The magnitude might seem a little underwhelming; however, it can be argued that the relationship between age and Body Mass Index is non-linear⁸, which means that the Pearson correlation coefficient will only pick up part of this relationship. As a next step, I plan to study the feasibility of a regression model to look a little more in depth at the issue. The results are in line with my previous hypothesis.

Careful consideration lead me to abandon 2.1⁹, at least in its original form. The reasoning is pretty straightforward: the non-linear relationship between *Age*, *Height*, and *BMI*, as well as the *Sex* differences, makes any naïve ranking not that meaningful. I am working on a better measure and definition of "fitness performance" to rank countries.

I provide a preliminary answer to question 2.2¹⁰ by looking at the time plots of BMI. Overall, an increase in average winners' BMI can be observed, more pronounced for men than for women. Pseudo-seasonality seems an issue here as well, so the next step will be to properly account for the difference between Summer and Winter Olympics. Apart from the decision to focus on winners rather than athletes in general, these

⁵"Has the age profile of Olympic Games winners changed over time?"

⁶"Is there a correlation between age and BMI?"

⁷I used the R's package 'Hmisc' `rcorr()` function for this purpose.

⁸Before the 20s, *Height* usually increases proportionately more than *Weight*, especially for professional-level athletes, whose nutrition is usually carefully controlled. Professionals often refer to percentiles of the appropriate age-related distributions.

⁹"Which are the top 3/5/10 countries by athletes' BMI (Male/Female)?"

¹⁰"How has average BMI changed over the years?"

results contradict my previous hypothesis. As noted before, BMI is quite a complex measure, and “increase in BMI” by itself does not mean “reduction in fitness”, since there are age- and sex-specific ranges for which BMI signals a healthy body, not point measures.

Regarding question 3.1¹¹, the descriptive statistics reported in the table suggest a potentially systematic difference: Winter Games winners tend to have higher BMIs (both male and females). The next step is to formally test this difference by performing a difference-in-means (between the two groups) statistical test.

Finally, question 3.3¹² has been temporarily set aside due to a lack of time. I am working on finding a simple and effective way to automatically categorize sports between “Team” and “Single” disciplines. If this approach does not prove to be fruitful, I will likely reconsider answering this question, since it might prove to be quite time-consuming.

Addendum

Most of the data retrieval and data management part was done in SQL (MySQL Workbench), while the rest, especially as far as graphs and more advanced tools were required, was done in RStudio. The present document has been created with R Markdown. All rights reserved.

¹¹“Is average BMI different between Summer and Winter Olympic Games?”

¹²“Is average BMI different between team and single disciplines?”