SQL for Data Science Capstone - Milestone 3

Francesco Aldo Tucci, 2022-01-11

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 profile[1].

[1]Some 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 group[2].

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

[2]The 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

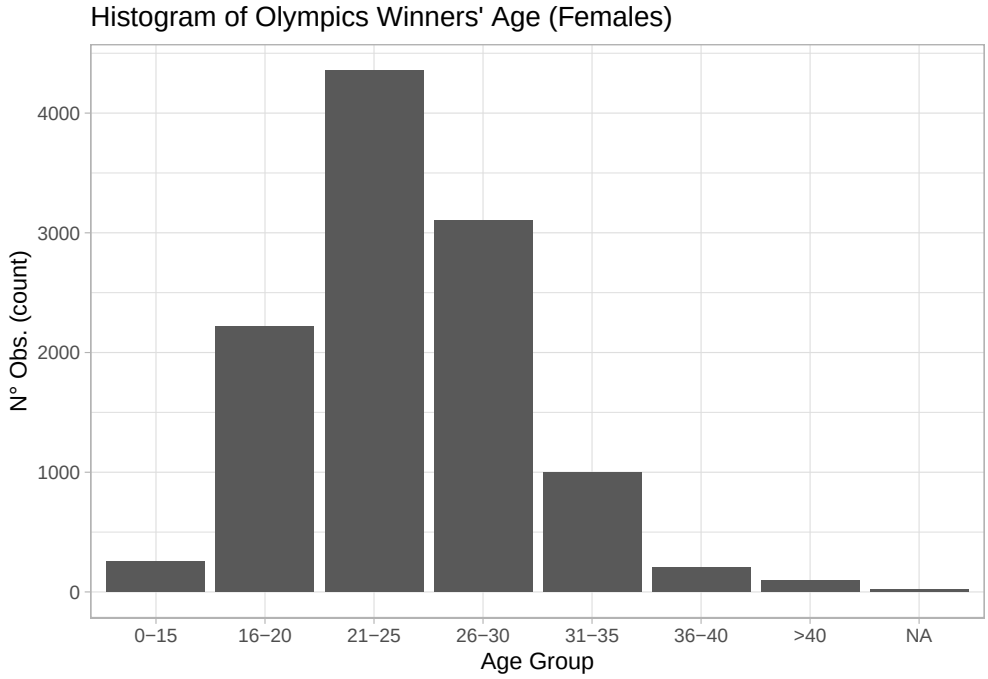
Table 8 - Descriptive Statistics (Winter Winners, Females)

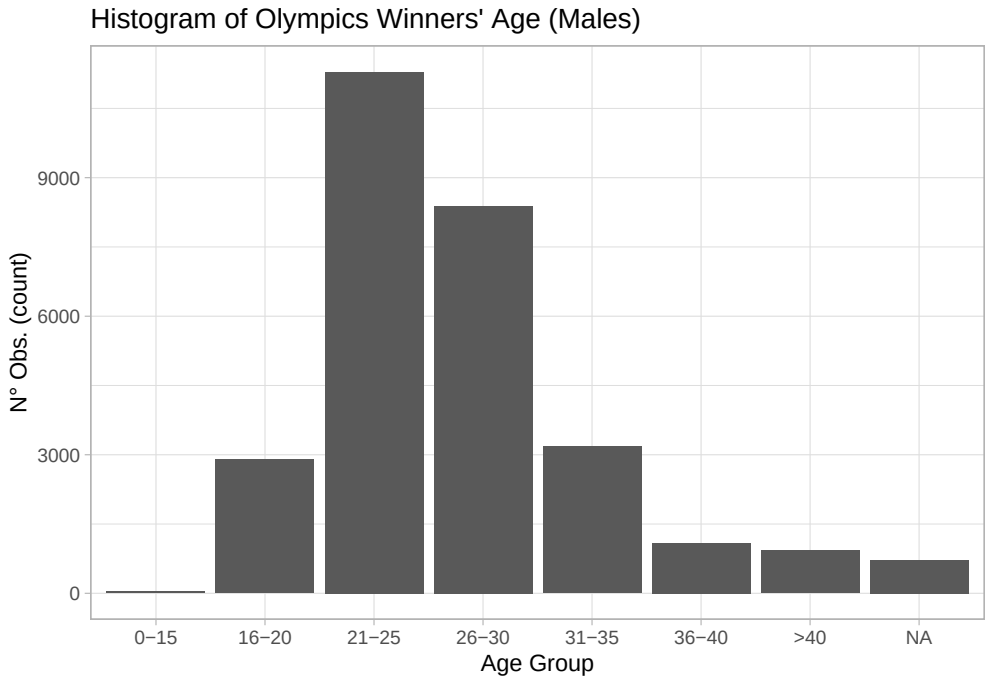
Winter win F	Age	Height (cm)	Weight (kg)	BMI
Min.	13	145.0	36.0	16.66
1st Q.	22	163.0	56.0	20.45
Median	25	167.0	61.0	21.76
Mean	25.19	167.3	61.38	21.85
3rd Q.	28	172.0	66.0	23.05
Max.	46	187.0	95.0	32.87
Range	33	42.0	59.0	16.21
IQR	6	9.0	10.0	2.60
Std. Dev.	4.8249	6.2086	8.0023	2.0625
N° Obs.	1811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

Table 9 - Descriptive Statistics (Winter Winners, Males)

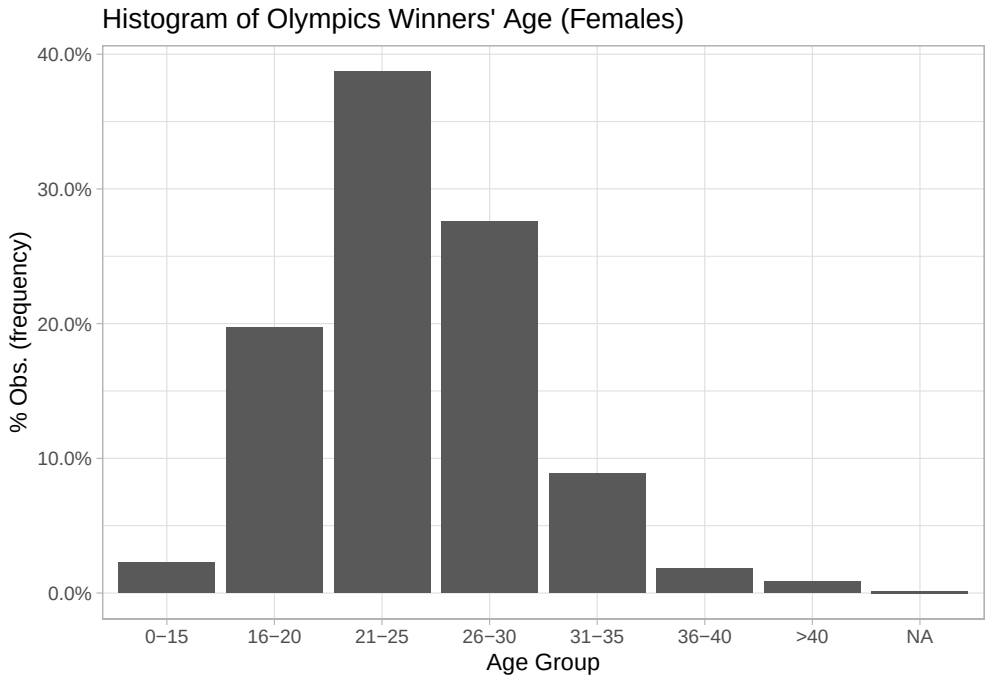
Winter win M	Age	Height (cm)	Weight (kg)	BMI
Min.	14	152.0	50.0	17.24
1st Q.	23	176.0	71.0	22.54
Median	26	180.0	79.0	24.30
Mean	26.47	179.9	78.88	24.26
3rd Q.	29	184.0	86.0	25.93
Max.	58	201.0	130.0	34.07
Range	44	49.0	80.0	16.83
IQR	6	8.0	16.0	3.39
Std. Dev.	4.8241	6.4138	10.9489	2.5074
N° Obs.	3884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

Then, 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)
40.0%
30.0%
20.0%
10.0%
0.0%
% Obs. (frequency)
0−15
16−20
21−25
26−30
31−35
36−40
>40
NA
Age Group

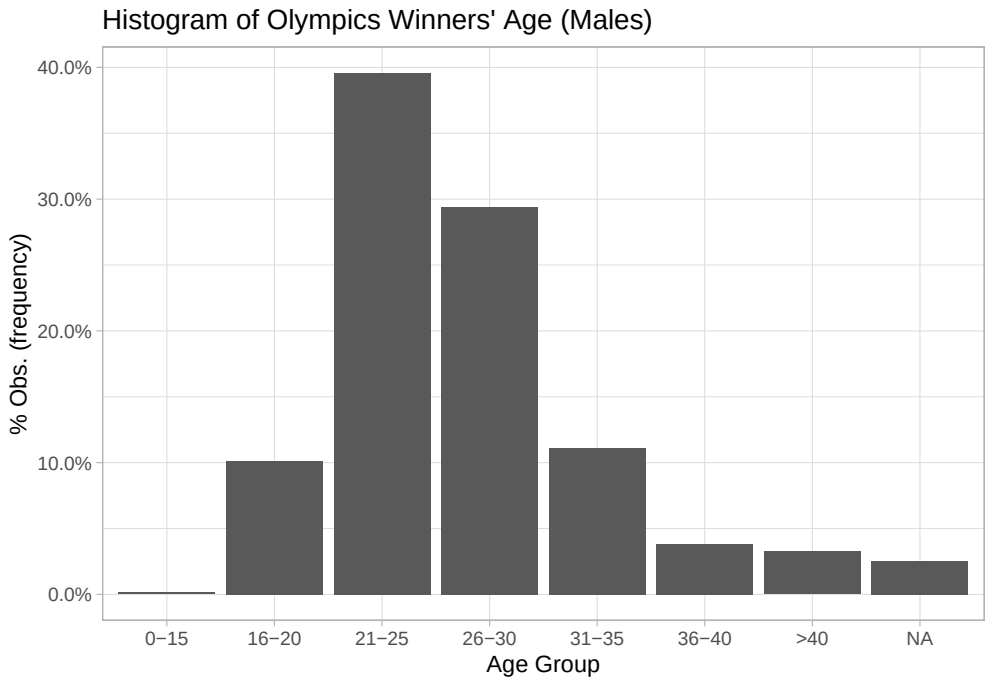
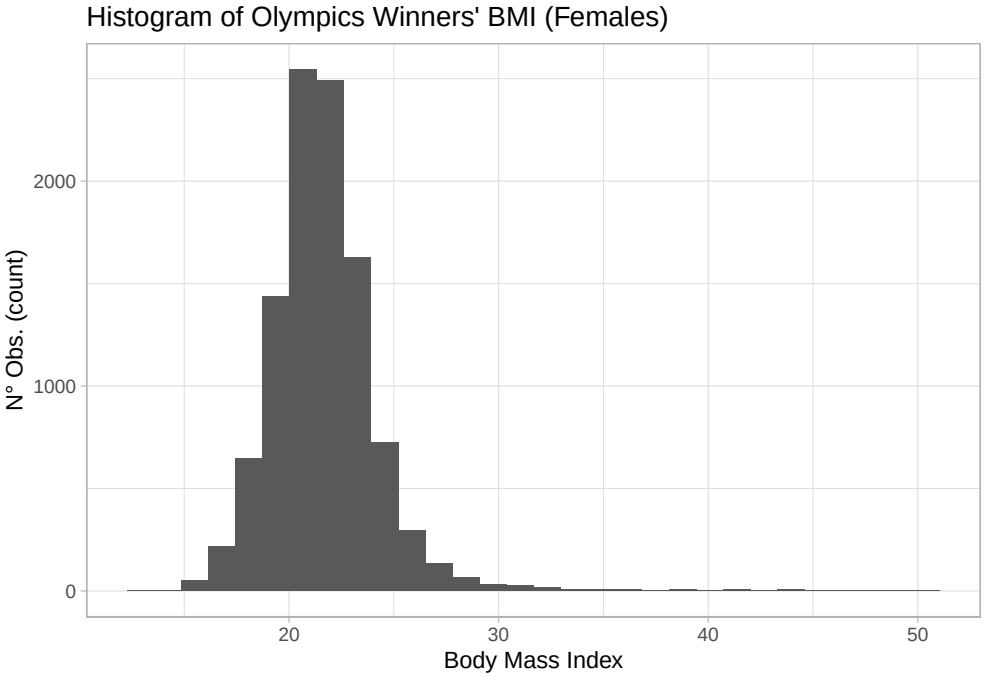
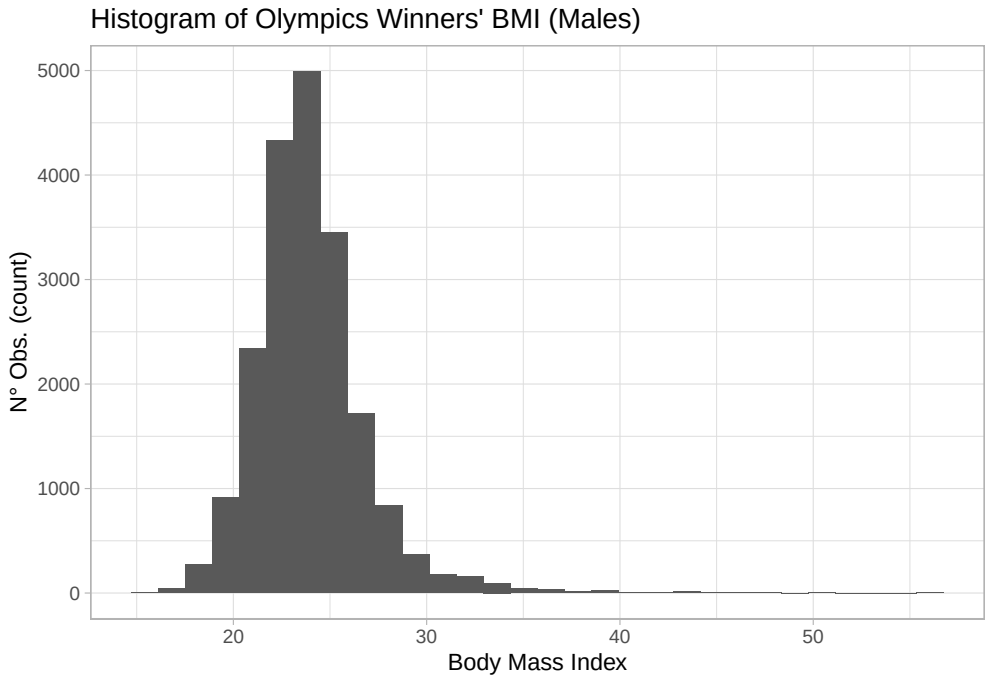

Histogram of Olympics Winners' Age (Males)
40.0%
30.0%
20.0%
10.0%
0.0%
% Obs. (frequency)
0−15
16−20
21−25
26−30
31−35
36−40
>40
NA
Age Group

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%





Now, I present a correlation matrix. Each entry represents the Pearson correlation coefficient, computed using pairwise complete observations[3], rounded to two decimal points, together with a symbol for statistical significance[4].

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

[3]While not reported, all correlation pairs use a sample size of at least 30,000 observations out of the 39,783 total.
[4]The following convention has been adopted: ^ 10%, * 5%, ** 1%, and *** 0.1% or less.

I asked myself how to sort of *rank* countries in terms of the BMI measure. BMI is a difficult index to use for such an exercise, since the "best" values are those at the center of the age-appropriate distribution, rather than the values of a monotonically increasing (decreasing) series. After consulting with some health professionals, I decided that it would be nonsensical to consider simple measures (different from population distribution percentiles), and that classifying countries "in reverse", based on the highest average BMIs, would be enough to satisfy my curiosity. To preserve some meaning, I only considered countries for which I have at least 20 observations[5] when considering the single Games, and at least 100 overall. The "top ten" results are summarized in the tables below.

Table 12 - Highest Average BMI by Country and Edition, Males

	Country	Edition	Avg. BMI	N. Obs.
1	Singapore	1956 Summer	27.8	49
2	Syria	1980 Summer	27.5	68
3	Georgia	2016 Summer	27.4	31
4	Guatemala	1952 Summer	27.3	33
5	Iceland	1992 Summer	27.1	26
6	Estonia	2012 Summer	26.7	25
7	Austria	1952 Winter	26.4	58
8	Iran	2012 Summer	26.3	46
9	Armenia	2016 Summer	26.1	24
10	Fiji	2016 Summer	26.1	36

Table 13 - Highest Average BMI by Country and Edition, Females

	Country	Edition	Avg. BMI	N. Obs.
1	Romania	1952 Summer	26.5	61
2	canada	1964 Winter	23.9	26
3	Camerun	2012 Summer	23.8	22
4	Canada	1960 Winter	23.6	24
5	Sweden	2006 Winter	23.6	73
6	URSS (Russia)	1952 Summer	23.6	97
7	Nigeria	2008 Summer	23.4	47
8	Nigeria	2016 Summer	23.4	28
9	Austria	2002 Winter	23.3	32
10	Cuba	1992 Summer	23.2	57

[5]If I were to conduct this type of analysis in a professional context, I would formally address (potential/likely) sample selection, rather than using such a convenient workaround! I chose 20 observations since they are enough to give us some confidence that I am not excluding smaller countries or Winter games, which see fewer athletes compete overall, *a priori*. I will consider using the median BMI, instead of the average, in the future.

Table 14 - Highest Average BMI Overall, Males

	Country	Avg. BMI	N. Obs.
1	Georgia	25.6	214
2	Latvia	24.9	596
3	Syria	24.9	138
4	Croatia	24.9	640
5	Armenia	24.8	183
6	Monaco	24.7	135
7	Fiji	24.7	165
8	Uzbekistan	24.5	344
9	Greece	24.4	1356
10	Egypt	24.4	1109

Table 15 - Highest Average BMI Overall, Females

	Country	Avg. BMI	N. Obs.
1	Angola	22.8	129
2	Egypt	22.4	181
3	Cuba	22.3	588
4	Nigeria	22.3	309
5	Venezuela	22.2	221
6	New Zealand	22.2	773
7	Chile	22.0	109
8	Serbia	21.8	139
9	URSS (Russia)	21.7	1590
10	Mongolia	21.7	149

Formally Testing for Difference In Means

Here I report the results[6] of the Welch t-tests for unequal variances, used so as to formally test the Null Hypothesis of whether two populations have the same mean.

Test	p-value	Significant at	H_0 Action
Overall M vs F	<0.00001	> 0.01%	Reject
Summer vs Winter (M)	<0.00001	> 0.01%	Reject
Summer vs Winter (F)	<0.00001	> 0.01%	Reject
Winners vs Not (M)	<0.00001	> 0.01%	Reject
Winners vs Not (F)	<0.00001	> 0.01%	Reject
Winners, S. vs W. (M)	<0.00001	> 0.01%	Reject
Winners, S. vs W. (F)	<0.00001	> 0.01%	Reject

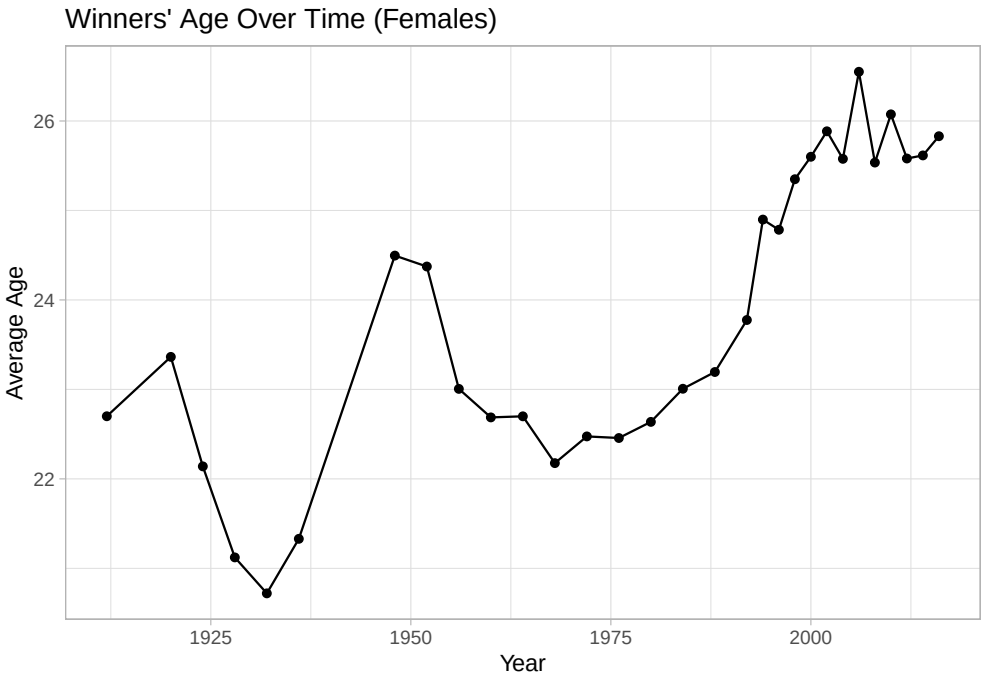
Next to do: apply Bonferroni and BH correction to p-values to account for multiple hypotheses testing.

[6]Reported results are not an error: all but one p-values are <2e-16; the other is 9.245e-13.

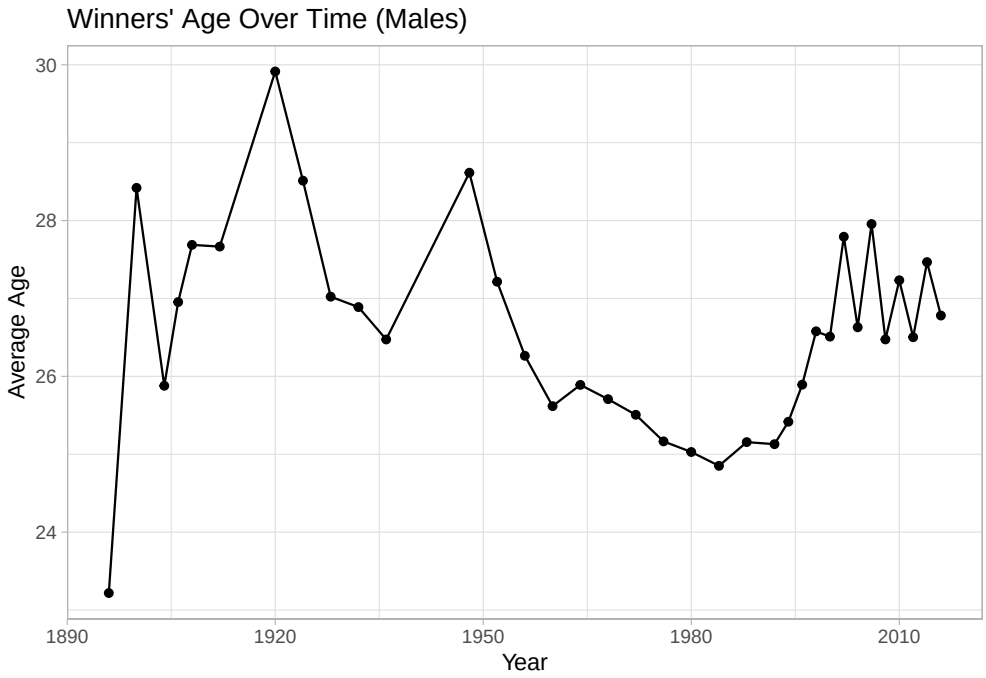
Linear Regression

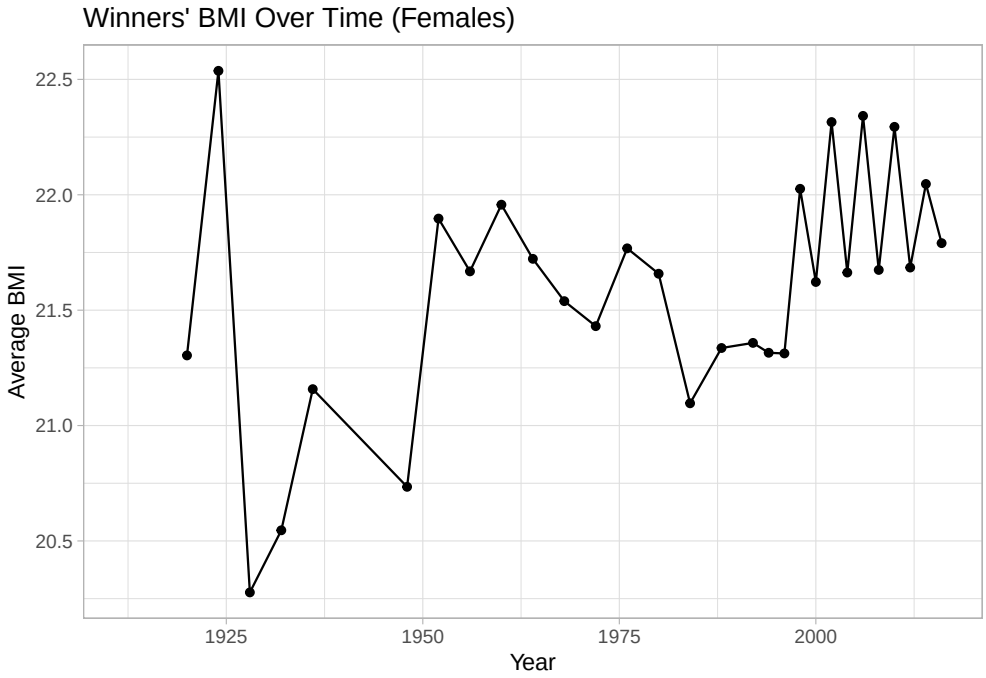
Below I report the summary output table for the principal linear regression models estimated through OLS. I included the estimated coefficients, standard errors, significance level symbols[7], adjusted R-squared and number of observations used in each regression. I will talk more about them in the Summary of Findings section.

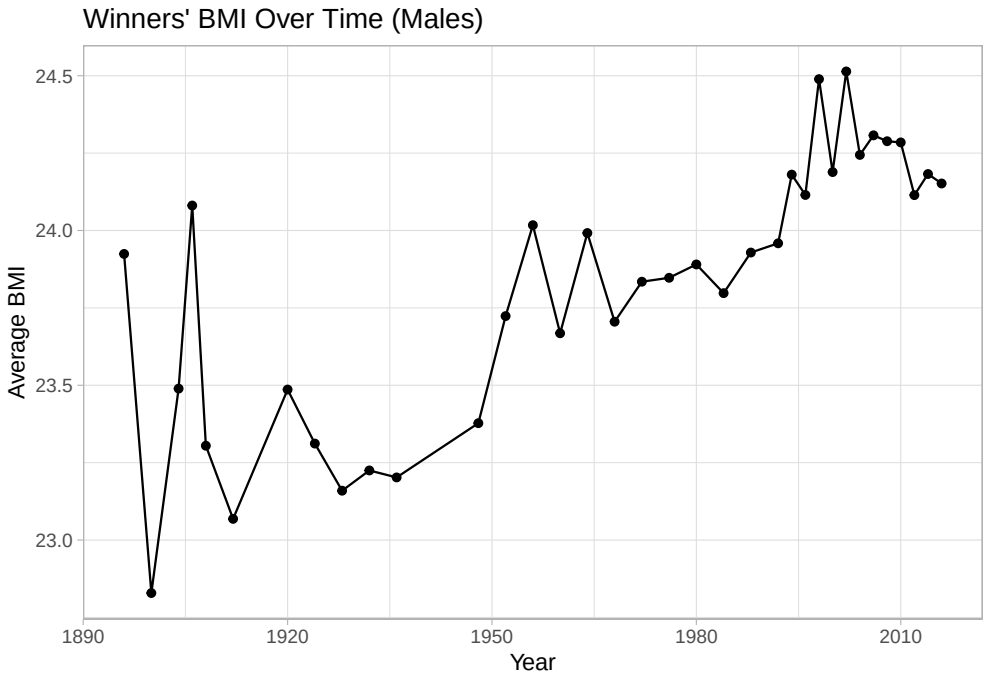
Model	(1)	(2)	(3)	(4)
(Intercept)	21.1932***	17.1021***	9.5889***	10.8048***
	(0.0104)	(0.0796)	(0.1242)	(0.0493)
Male dummy	2.34908***	2.1619***	1.6353***	-0.4385***
	(0.01269)	(0.0127)	(0.0142)	(0.0088)
Age		0.2310***	0.1760***	-0.0315***
		(0.0058)	(0.0057)	(0.0035)
AgeSq		-0.0026***	-0.0017***	0.0010***
		(0.0001)	(0.0001)	(0.00006)
Winter dummy		0.3461***	0.3884***	0.3100***
		(0.0148)	(0.0146)	(0.0090)
Medal dummy		0.4493***	0.3166***	-0.1847***
		(0.0165)	(0.0163)	(0.0100)
Height			0.0495***	
			(0.0006)	
Weight				0.1750***
				(0.0003)
Adj. R-squared	0.1422	0.1765	0.2001	0.6984
N° Obs.	206853	206165	206165	206165

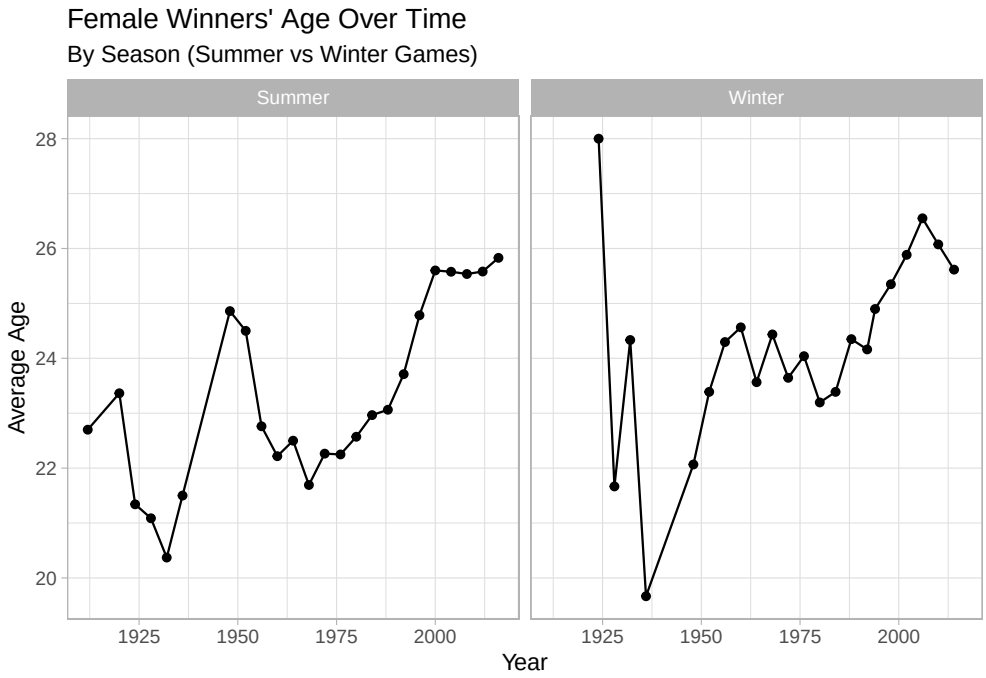


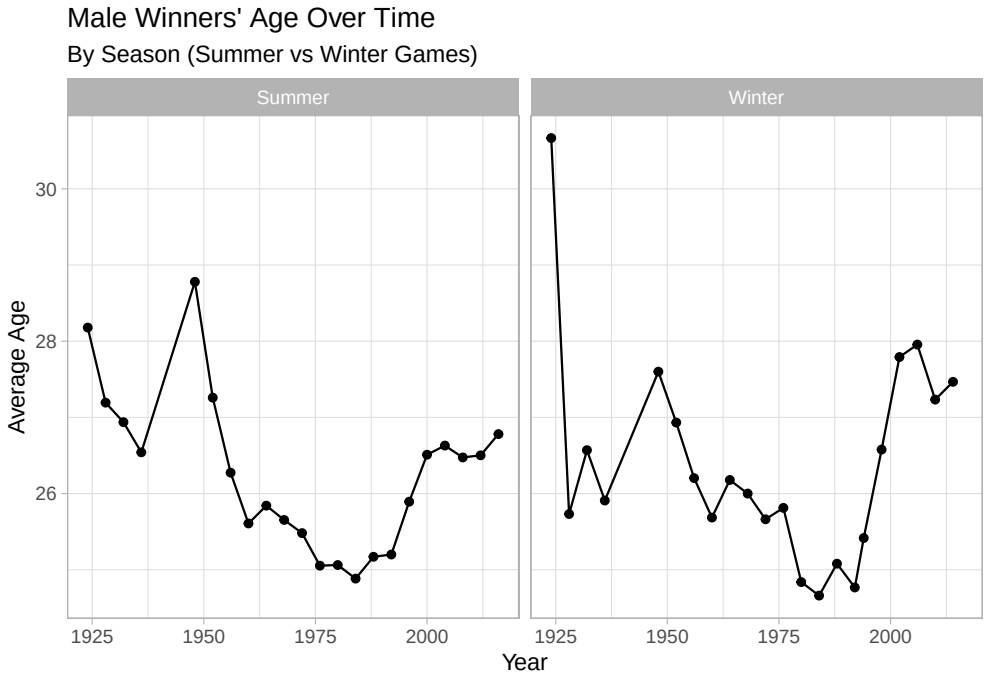
[7]Statistical significance conventions: *** <0.001%, ** 0.01%, * 0.05%, ˆ 0.1%.

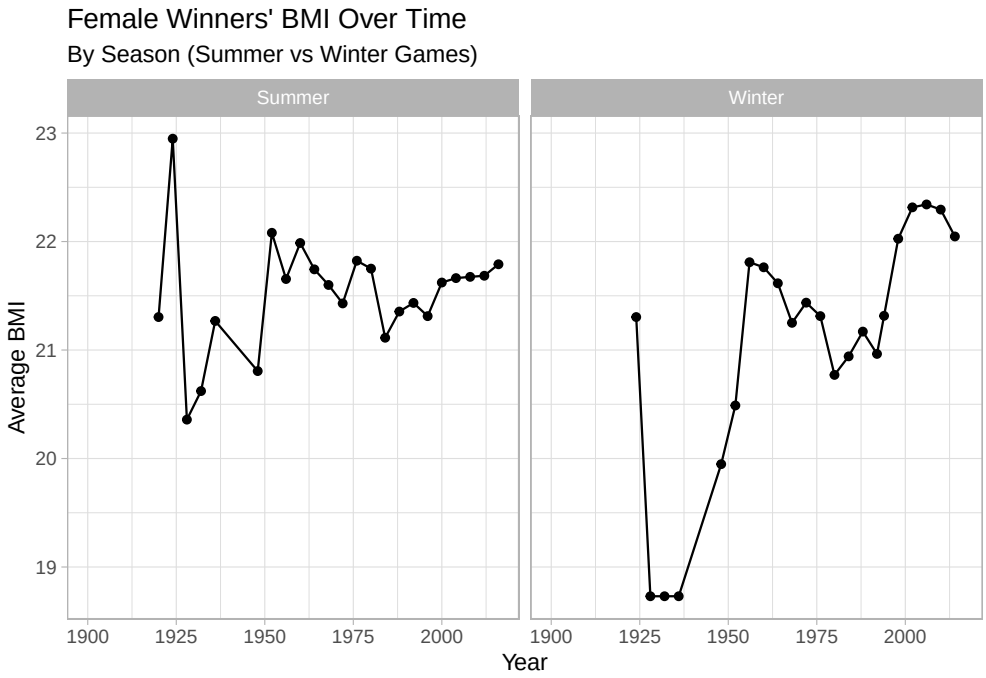

Winners' Age Over Time (Males)
Average Age
30
28
26
24
1890
1920
1950
1980
2010
Year

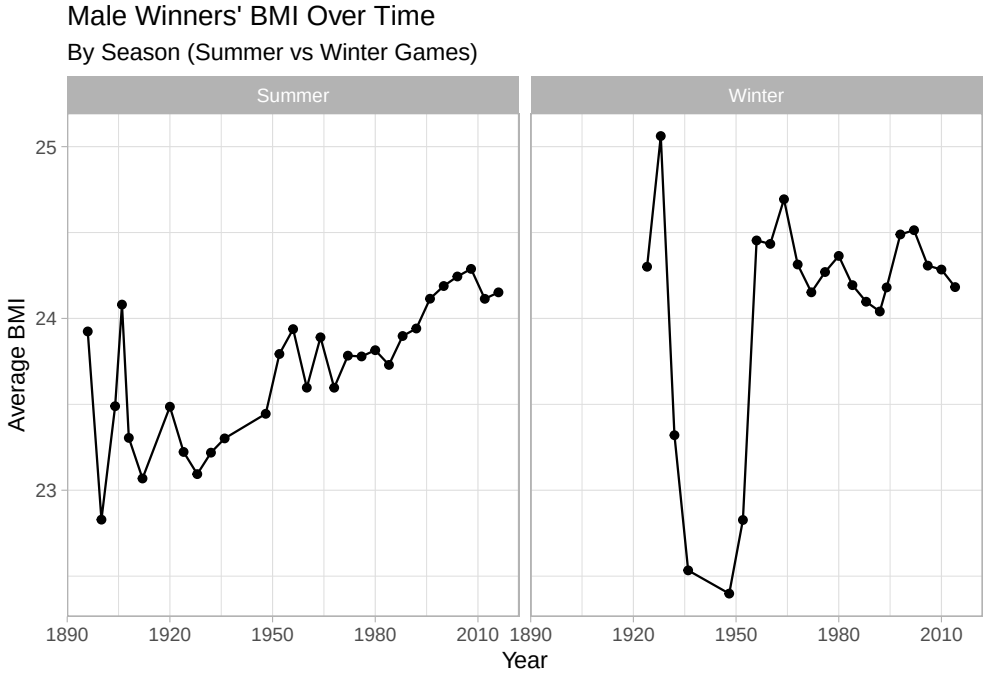

Winners' BMI Over Time (Females)
Average BMI
22.5
22.0
21.5
21.0
20.5
1925
1950
1975
2000
Year


Winners' BMI Over Time (Males)
Average BMI
24.5
24.0
23.5
23.0
1890
1920
1950
1980
2010
Year


Female Winners' Age Over Time
By Season (Summer vs Winter Games)
Summer
Winter
Average Age
28
26
24
22
20
1925
1950
1975
2000
Year


Male Winners' Age Over Time
By Season (Summer vs Winter Games)
Summer
Winter
Average Age
30
28
26
1925
1950
1975
2000
Year


Female Winners' BMI Over Time
By Season (Summer vs Winter Games)
Summer
Winter
Average BMI
23
22
21
20
19
1900
1925
1950
1975
2000
Year


Male Winners' BMI Over Time
By Season (Summer vs Winter Games)
Summer
Winter
Average BMI
25
24
23
1890
1920
1950
1980
2010
Year

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 1.1[8] 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 was 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 and 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. The overall trend (a decline from 1950s until around 1975 for females and 1980 for males, followed by un upward trend that seems to continue up to the most recent editions) is still clearly present in the graphs faceted by Season.

To answer 1.2[9], and actually go a little bit beyond that, a correlation matrix is shown. The statistical significance of the computed correlations has been formally tested[10]. 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, which makes it significant at 10% only). 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-linear[11], which means that the Pearson correlation coefficient will only pick up part of this relationship. Other notable correlations are between the *Male* dummy and *Height* and *Weight* (overall, males tend to be taller and weigh more than females, as it is well known), which in turn is reflected in a clear and statistically significant linear correlation between the athlete being male and *BMI*. The results are in line with my previous hypothesis.

As argued before, careful consideration lead me to abandon 2.1[12] 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 satisfied my curiosity (and hopefully the reader's too) by looking at the "worst" countries, i.e. teh countries that showed the highest average BMIs. As hinted previously, this "worst" ranking is relative and just for flavor: it does not really mean that the countries

[8] "Has the age profile of Olympic Games winners changed over time?"

[9] "Is there a correlation between age and BMI?"

[10] I used the R's package 'Hmisc' rcorr() function for this purpose.

[11] Before 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.

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

identified have athletes with worse physical condition or fitness status compared to the others. The tables with the rankings are pretty much self-explanatory.

I provide a preliminary answer to question 2.2[13] 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 is 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 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. Accounting for the differences between Summer and Olympic Games show a quite clear upward trend in BMI over time for Summer games, while the Winter editions' picture is less clear-cut, actually showing a decrease lately (from the peak for females, from a local maximum for males).

Regarding question 3.1[14], 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 has been to formally test this difference by performing a difference-in-means (between the two groups) statistical test. The results are pretty clear-cut, and while I plan to properly perform Bonferroni and Benjamini-Hochberg correction to account for the (arbitrary choice of) number of Null Hypotheses tested, the p-values are small enough to make it unlikely to see any dramatic change in my conclusions. First, I find reassurance of the statistically significant difference in average BMI between males and females. As argued for the rankings, a proper interpretation of these differences might lead us to unfruitful endings, but it is interesting to note that winners of Olympic medals show systematic differences in BMI over the rest of the athletes, and that this is true for the grouping by *Season* (i.e. type of sports and disciplines!) as well.

Finally, question 3.2[15] has been definitely set aside due to a lack of time. I will work on finding a simple and effective way to automatically categorize sports between "Team" and "Single" disciplines, but for the time being this will be left out of the final report.

I have other results ready that I plan to include in a "Miscellanea" and an "Appendix" sections respectively for the final report.

Meanwhile, let me offer a brief comment about the linear regression results. First, please note that this has been just an exercise intended to showcase how linear regression can be used to gain further insights about a novel dataset. Second, many of the issues will be resolved or at least properly addressed for the final report. For now, let's proceed as the usual OLS assumptions[16] were met. Model (1), the simple (univariate) regression of *BMI* on a *Male* dummy, quantifies the statistically significant linear relationship betweeb "being male" and BMI previously identified through the Pearson correlation coefficient. All the other regressors added in model (2) are statistically significant as well. While (again, repetita iuvant) an increase in BMI does not always have a direct or clear-cut interpretation, it is interesting to highlight how, all other things equal, winners tend to have higher BMIs compared to other athletes. The variables *Height* and *Weight*, which directly enter the formula for the BMI metric, are entered in turn to study how they affect such a regression exercise. There is a vast increase in the adjusted R-squared, which makes the model much more powerful at predicting the BMI values. However, while *Height* is still somewhat fine, since it enters the BMI formula nonlinearly, the inclusion of *Weight* is different. Since BMI can be defined as weight times a factor which depends on height, *Weight* alone explains most of the variation in *BMI*. The model acquires very strong predictive power (as signaled by a 0.9802 adjusted R-squared), but the estimates of the other coefficients lose (further) reliability. An extreme case will be included in the Appendix for illustration purposes.

[13] "How has average BMI changed over the years?"

[14] "Is average BMI different between Summer and Winter Olympic Games?"

[15] "Is average BMI different between team and single disciplines?"

[16] For reference, say assumptions MLR1-MLR4 from Wooldridge (2015), *Introductory Econometrics: A Modern Approach*, Cengage learning.

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.