Olympic Medals: Matter of Nerves

Dr Preet Deep Singh

21/08/2021

Abstract

Using a data of over 120 years of Olympic history, I find that the mean age of medal winners is higher than that of non-medal winners. This relationship is more pronounced in the case of Gold medalists. I hypothesize that Age is a proxy for experience of Olympic Games and use show that the probability of getting a medal in the second appearance is higher than that in the first one. This could be due to experience, better handling of pressure, and better preparation. Our results are robust to change in models, controls, proxies, and years.

1 Introduction

I look at the age of athletes that have competed in Olympics and see that there is a difference between the age of those who got a medal and those who did not. This difference is statistically significant post 1950. We examine this further and see that the difference is more pronounced in case of Gold medalists and non-medalists. While the difference is not much, in my opinion higher age would lead to slower reflexes and worse performance. I therefore explore whether there is a component of experience in this. I use the number of appearances in Olympics as a measure of experience. I hypothesize that more exposure to the big stage would allow players to hold their nerves and therefore improve their performance. There are a host of psychological factors at play here as well. There is fear of repeated non-performance following a medal-less performance in the first time. I find that athletes are more likely to get a medal in their subsequent appearances. Out results are consistent in direction when we control for Years, Sports, Age, Gender, Weight and Height. Apart from establishing the role of experince, my results offer a ray of statistical hope to non-medalists from Tokyo Olympics.

Our analysis does not include data from Tokyo Olympics in 2021. Since we use more than a century worth of data, I expect my results to hold through.

2 Visualising Age and Success

I look at the age of all athletes that have competed at Olympics in the past 120 years. It has a mean of 25.56. The mean age of all athletes that won a medal over the same time period is 25.93. The mean age of all athletes who did not win a medal during the same period is 25.49. As evident, there is very little difference in the mean age of athletes who won and who didn't win. One reason for the same could be the quality of the competition where the difference is overcome by skill, training and practice.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      10      21      24      26      28      97    9474
```

2.1 Low Age

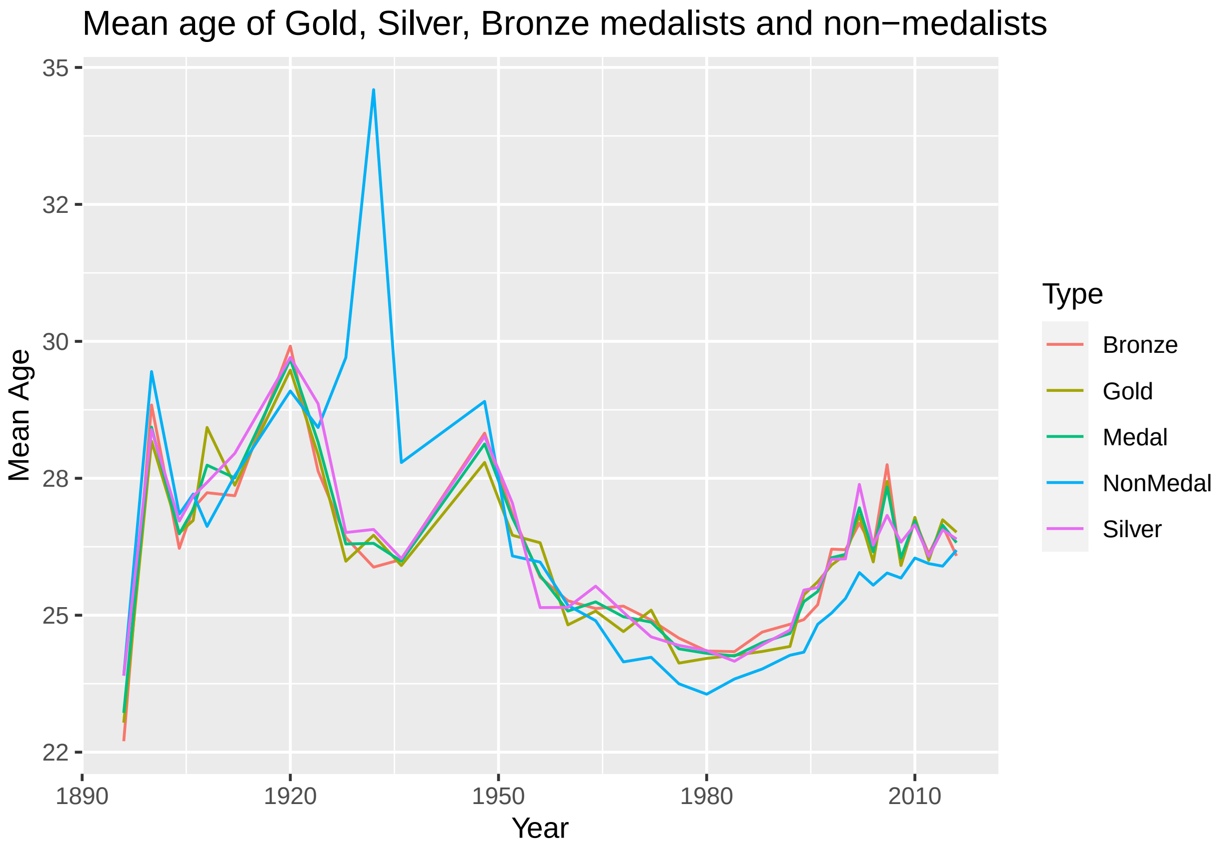
A total of 12508 appearances have been made at Olympics by athletes who were less than 18 years of age. Out of these 6712 have been after 1980. This number includes multiple participation by the same athlete as well. 5959 unique athletes have participated till date and those who competed after 1980 are 3136.

2.2 High Age

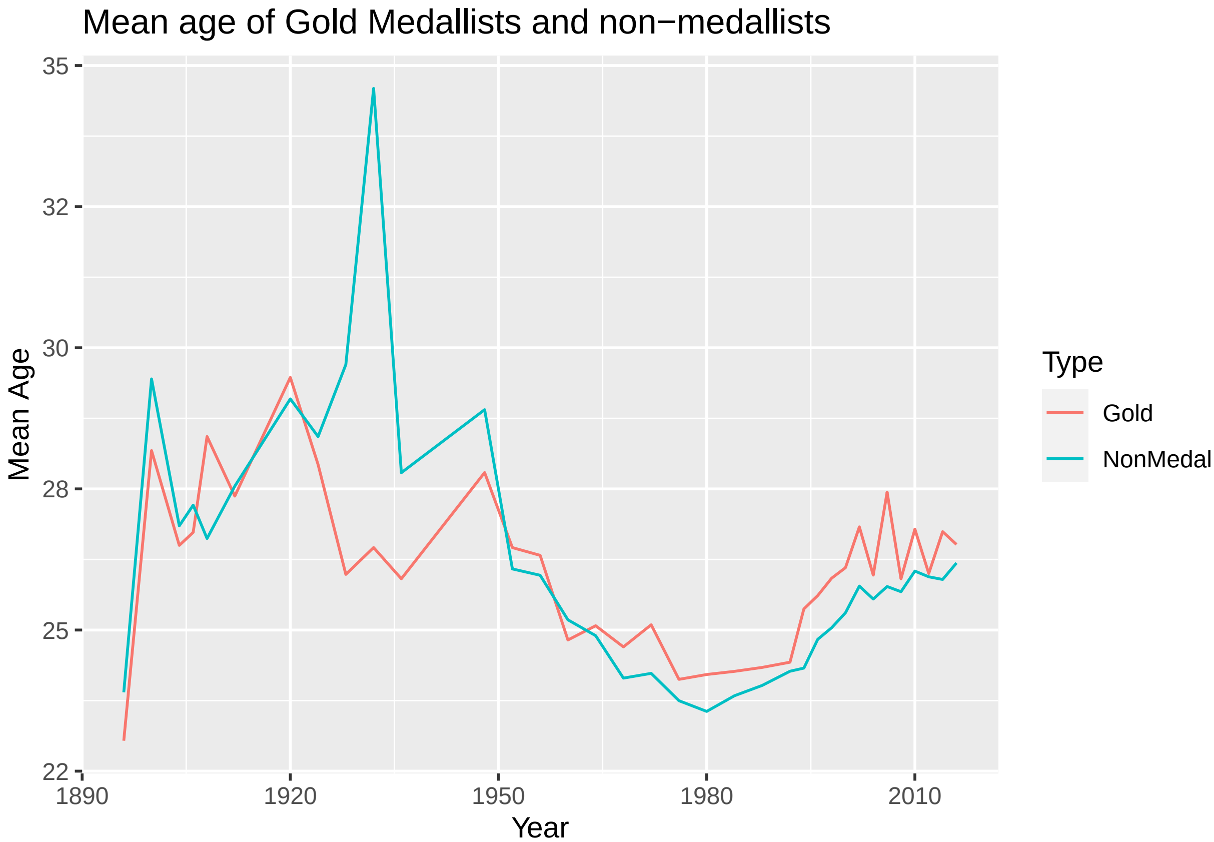
A total of 570 appearances have been made at Olympics by athletes who were more than 60 years of age. Out of these 26 have been after 1980. This number includes multiple participation by the same athlete as well. 235 unique athletes have participated till date and those who competed after 1980 are 16.

2.3 Pooled Plot

A pooled plot of the age of athletes over the years shows much more variation before 1950. The mean age of 1928 Olympics is abnormally high. While the age has not been constant in any two years, the mean age has remained between 25 and 28 in this century.



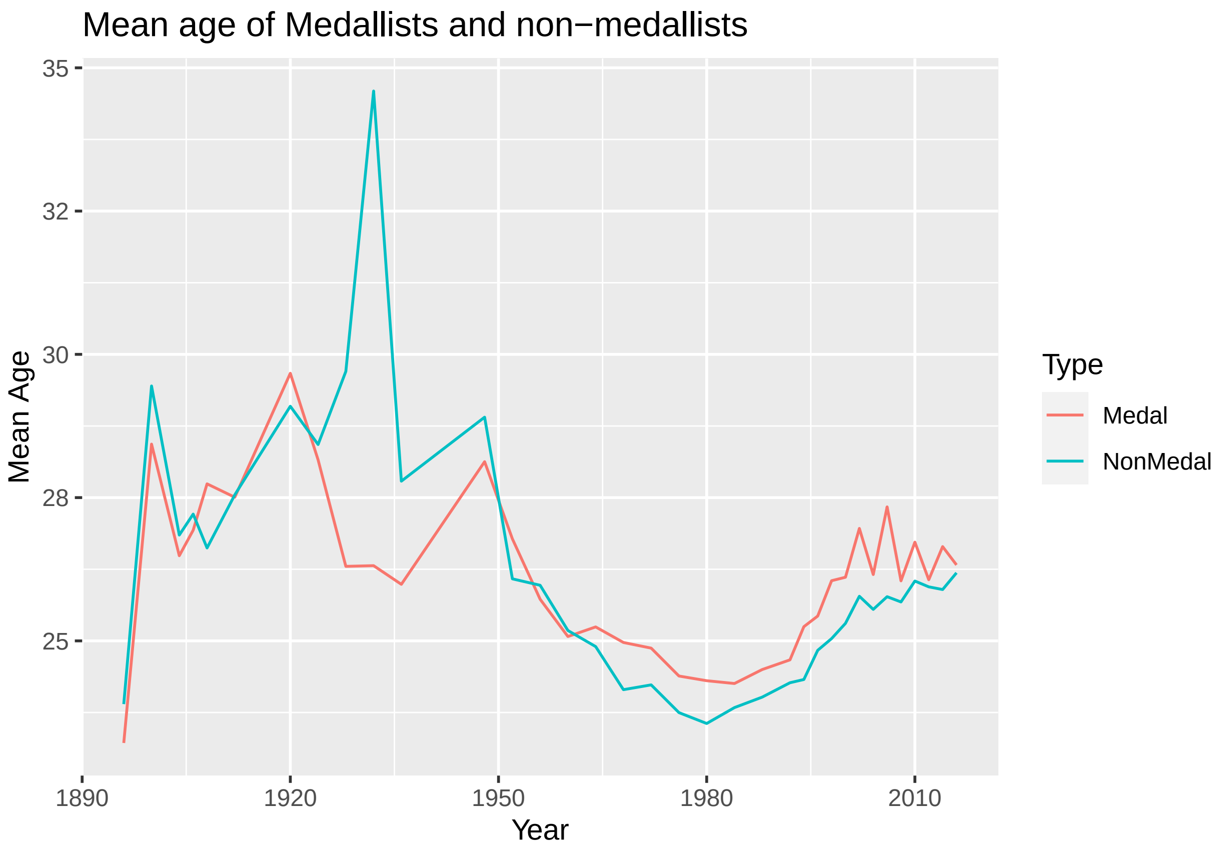
2.4 Gold Vs Non Medal



In order to look at the difference between the Gold Medal winners and the non-medalists, we plot their mean age over the years. Starting 1950, the mean age of Gold Medalists has been higher than that of non-medalists other than in one instance. This could be because age represents experience. Olympics is an important event and there is a fair amount of pressure on the participants. Age or experience might assist in dealing with that pressure. We conduct a Welch Two Sample t-test, the p-value of the difference in mean of age of gold medalists and non-medalists is 0.05 which is fairly robust. When we extend our sample to include data before 1950 we see that the p-value is 0.82. This could be because of the anomaly in 1928.

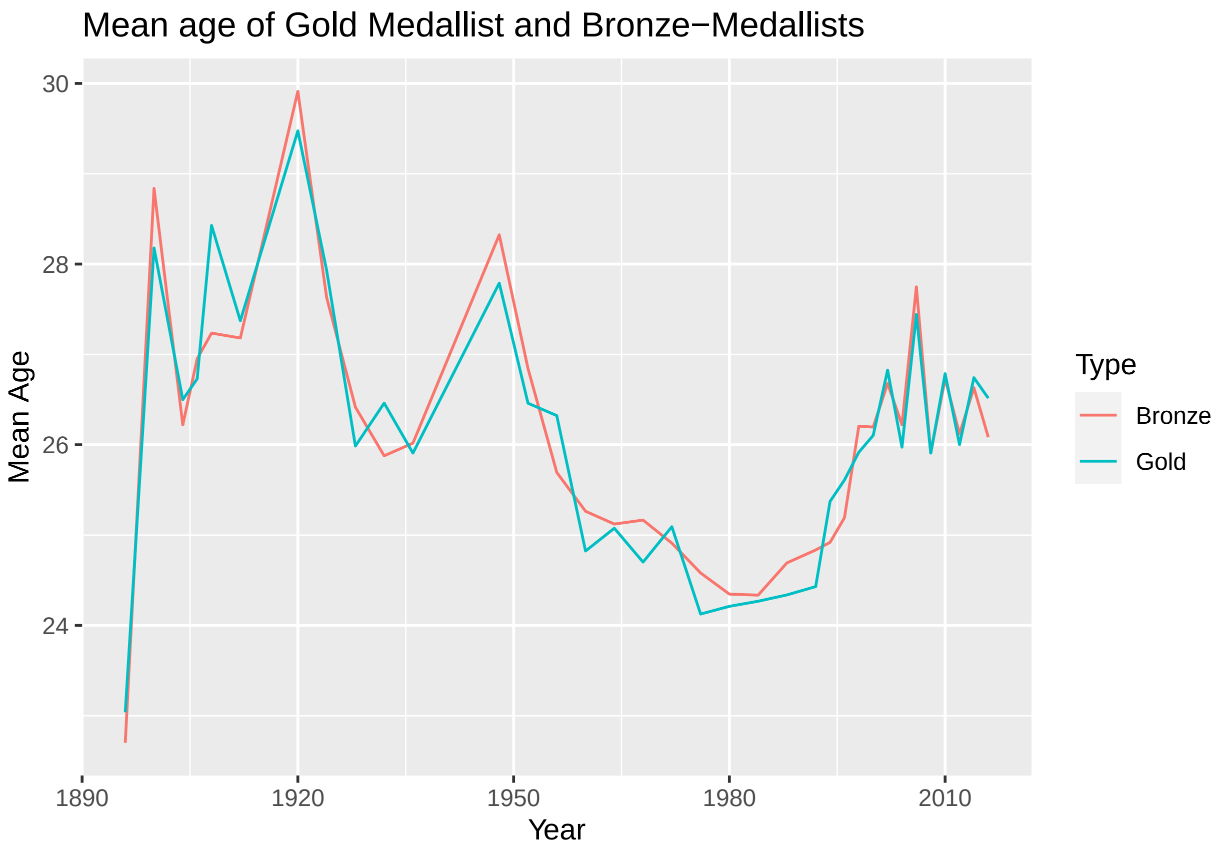
2.5 Medal Vs Nonmedal

I also look at the difference in mean age between medalists and non-medalists. Similar to above, I find that the mean ages are 25.65, the p value is 0.03 which indicates robustness. If we extend our sample to before 1950, we see that while mean ages are 26.14, the p value of 0.92 is not significant.



Mean age of Medallists and non−medallists
35
32
30
28
25
Mean Age
1890
1920
1950
1980
2010
Year
Type
Medal
NonMedal

2.6 Gold Vs Bronze



I look at values after 1950 for the mean age of Gold medal winners and Bronze medal winners. and find that the mean ages are 25.61 and 25.67 , the p value is 0.83 which implies that the observed difference is random. This points to the tough competition at the Olympics where the difference between medalists is not as much as the difference between a medalist and a non-medalist..

3 Regression

We use Logistic regression to model the dependent variable of Success defined as getting a medal. Our model is

We use a pooled regression to model success as defined by winning a medal, dependent on age, while controlling for height, weight, and gender. We see that the p value for Age is extremely significant at 2.28×10^{-14}, but the estimate of 0 is not materially significant. Change of an year in age would lead to 0.001% increase in the probability of a medal.

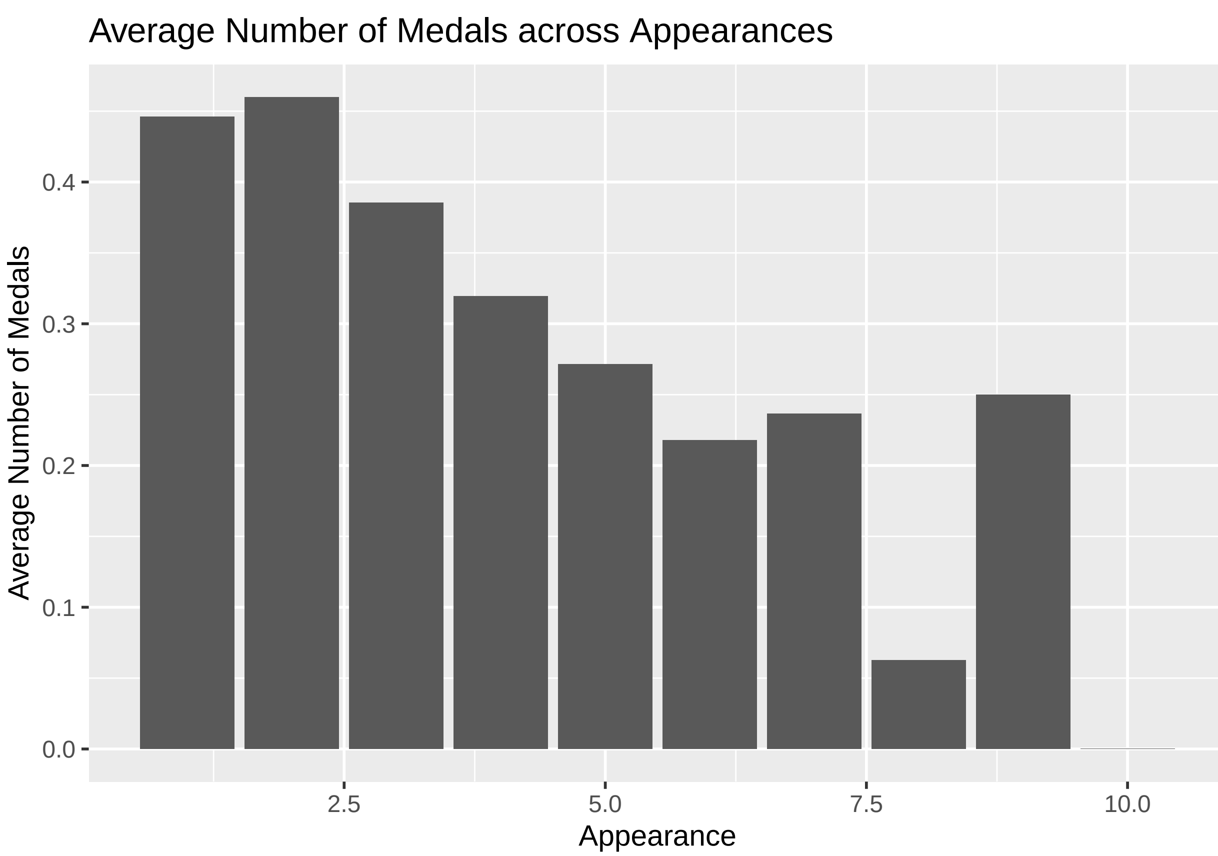
4 Medals Across Outings

I look at Olympians who have participated in multiple versions of Olympics across years and have bagged at least one medal. I filter this to see whether the medal tally in subsequent appearances. I hypothesize that it is hard to maintain supremacy in such a competitive tournament and athletes would not be able to win in subsequent versions. We see a list of 12994 athletes that have competed in multiple Olympics with the highest being Ian Millar Olympics. We do not consider breaks. For example if an athlete competed in two Olympics across 8 years, we still consider that as two participations irrespective of whether it was consecutive or not.

Table 1: GLM Regression Results for Medal on Age

	Medcod Whether Medal
Age	0.001***
	(0.0001)
Height	0.002***
	(0.0001)
Weight	0.002***
	(0.0001)
Constant	−0.400***
	(0.017)
Gender Controls	Yes
N	206,165
Log Likelihood	−76,561.000
Akaike Inf. Crit.	153,131.000
Notes:	***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

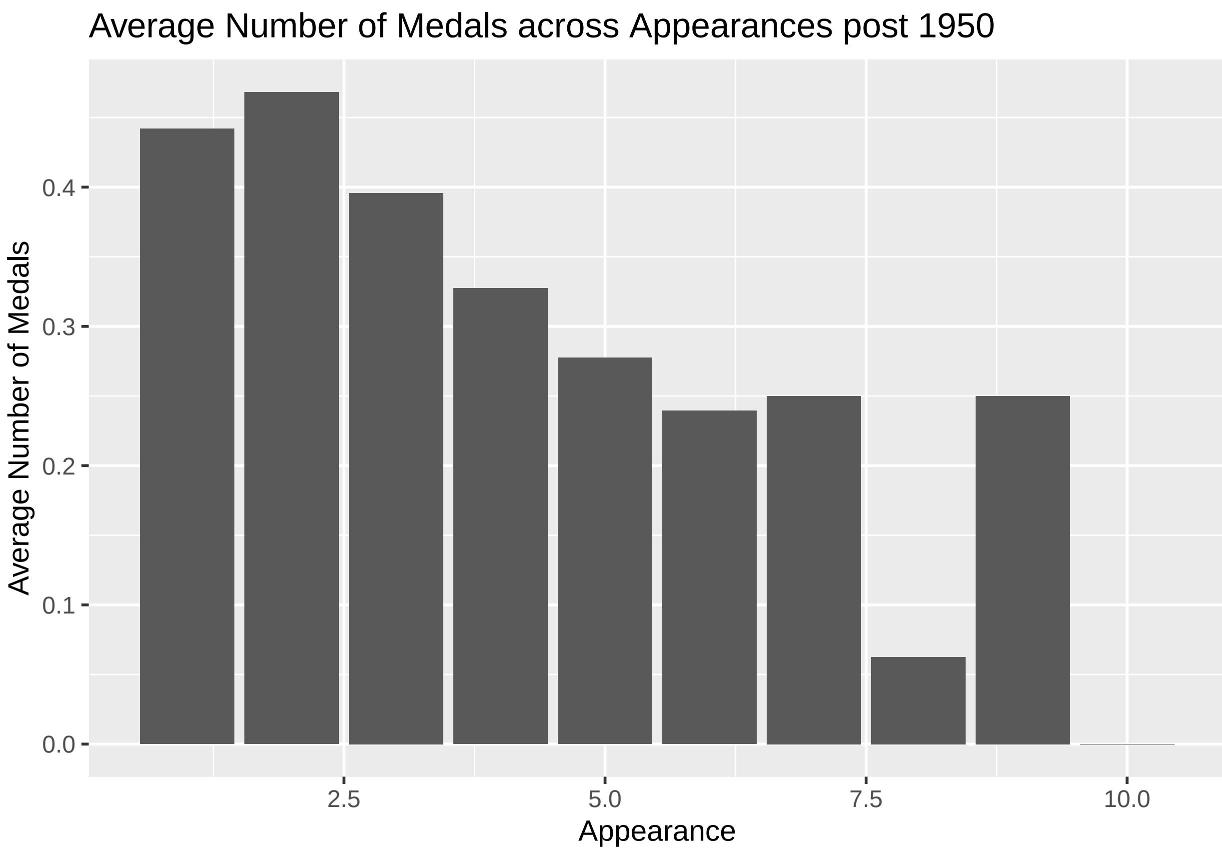
4.1 Pooled



If we look at the average number of medals (any color) for each iteration of appearance at the Olympics we find that the second outing is more likely to result in a medal than the first. I run a Welch Two Sample t-test on the medal victories in the first appearance versus a medal in the second appearance. The mean values are

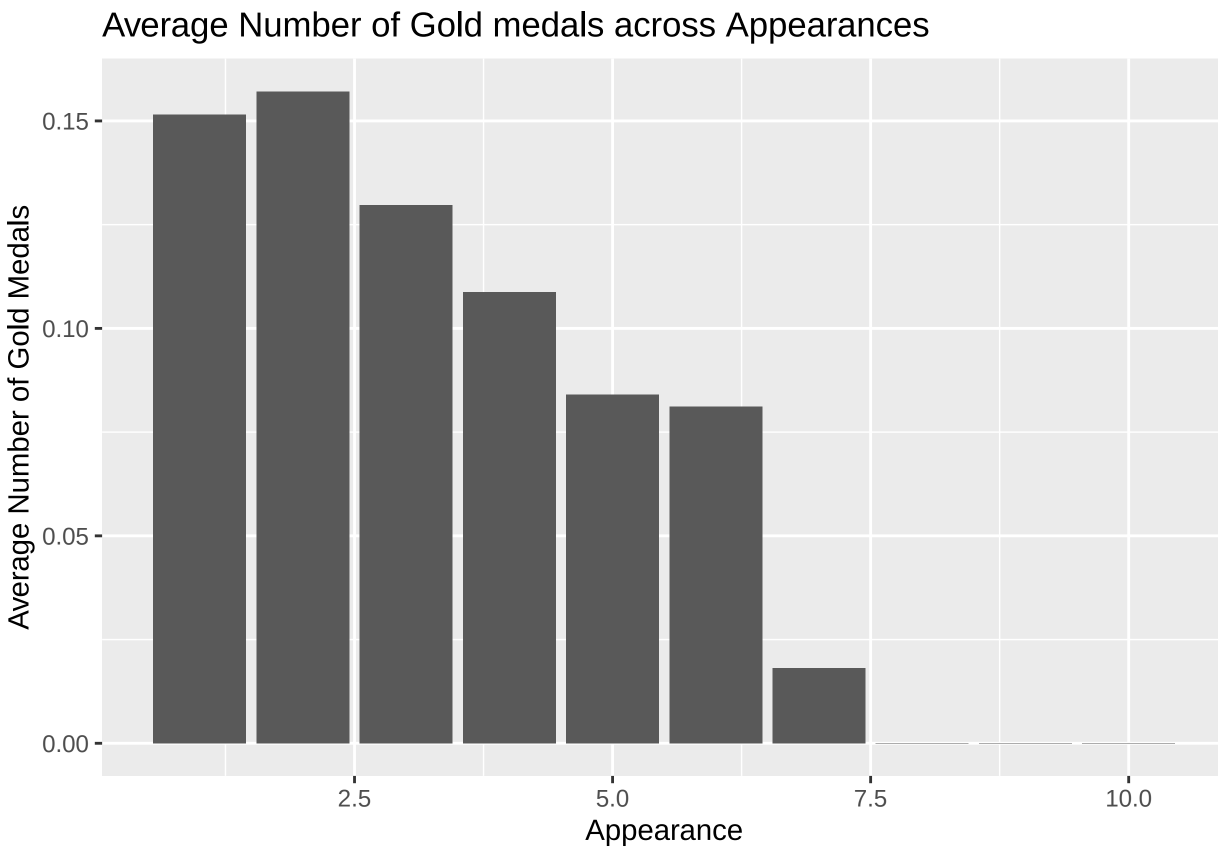
0.45 and 0.46 with a p value of 0 which indicates that this relationship is statistically highly significant.

4.2 Post 1950



The numbers from the pooled sample as well as recent sample are very similar and the average values differ from the third decimal point. Similarly when I look at results from after 1950, I see that the man number of medals in the first appearance are 0.44 as compared to a mean value of 0.47 in the second outing. The p-value for this difference is highly significant at 8.93×10^{-7}.

4.3 Gold Medals



While the maximum appearances of an athlete to have won a gold medal are 7, I am interested in looking at how gold medals stack across appearances. The graph shows that the average number of Gold Medals are higher in subsequent appearances, though by a very thin margin. I conduct a Welch Two Sample t-test and see that the p value of 0.11 is not significant, thereby signalling that the difference in success based on the iteration of appearance is not there. We look at the same analysis for a more recent period, that is after 1950 and find a similar p value.

We conclude that the difference in success defined as winning a gold medal is not influenced by whether it is the first of the second appearance but when defined as winning a medal, the sequence of the appearance matters.

It coincides with our reasoning on why higher age seems to be correlated with medals even when controlling for height, weight, and gender.

4.4 Gold Medals, post 1950

5 Regression on Appearances

ADD MODEL Our model is

$$Gold = \alpha + \beta_1 Appearances + \beta_2 Age + \beta_3 Height + \beta_4 Weight + \beta_5 Sex + \beta_6 Year + \beta_7 Sport + \epsilon$$

We see from the regression that Age is now negative and Count is positive. It is statistically significant at 0. This means that a younger person with prior experience is more likely to win a Gold medal. ###0.03

Table 2: GLM Regression Results for Gold Medal on Appearances

	Gold
Appearances	0.053***
	(0.001)
Age	−0.002***
	(0.0001)
Height	0.001***
	(0.0001)
Weight	0.0004***
	(0.0001)
Constant	0.120***
	(0.036)
Control for Sex	Yes
Control for Year	Yes
Control for Sport	Yes
N	206,165
Log Likelihood	27,137.000
Akaike Inf. Crit.	−54,083.000
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

Table 3: GLM Regression Results for Medal on Appearances

	Medcod Whether Medal
Appearances	0.150***
	(0.001)
Age	−0.005***
	(0.0002)
Height	0.002***
	(0.0001)
Weight	0.001***
	(0.0001)
Constant	0.110**
	(0.056)
Control for Sex	Yes
Control for Year	Yes
Control for Sport	Yes
N	206,165
Log Likelihood	−64,767.000
Akaike Inf. Crit.	129,725.000
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

We repeat this exercise using MedCod as the dependent variable to model whether winning any medal is dependent on the number of appearances and age controlling for height, weight, year, and sport. We see that the coefficient has gone up 0.16 and is statistically significant with a p-value of 0. For the sake of hygeine, we try multiple variations of the same exercise.

Table 4: GLM Regression Results for Medal across models on Appearances

	Medcod		
	Without Year and Sport	With Year	With Sport
	(1)	(2)	(3)
Count2	0.150^{***}	0.150^{***}	0.160^{***}
	(0.001)	(0.001)	(0.001)
Age	-0.005^{***}	-0.004^{***}	-0.006^{***}
	(0.0001)	(0.0001)	(0.0002)
Height	0.002^{***}	0.002^{***}	0.001^{***}
	(0.0001)	(0.0001)	(0.0001)
Weight	0.002^{***}	0.002^{***}	0.001^{***}
	(0.0001)	(0.0001)	(0.0001)
Constant	-0.360^{***}	0.013	-0.240^{***}
	(0.017)	(0.056)	(0.021)
Control for Sex	Yes	Yes	Yes
Control for Year	No	No	No
Control for Sport	No	No	No
N	206,165	206,165	206,165
Log Likelihood	−69,795.000	−68,585.000	−65,966.000
Akaike Inf. Crit.	139,601.000	137,251.000	132,054.000

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

We see that even if we change the model by dropping one or both of Sport and Year the coefficient and p-value are robust.

When we take only recent cases, that is Olympics after 1950, we see that the results are still robust. It seems this is a phenomenon for the ages!

6 ToDo

The source of this data is https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results/metadata. See if multi medal people have this too. Remove the Print Commands since we have Stargazer now.

- Limitations
- Titles for all graphs

7 Extra Analyses

THIS IS FROM THE SECTION REGRESSION

Table 5: GLM Regression Results for Medal after 1950 on Appearances

	Medcod Whether Medal
Appearances	0.160***
	(0.001)
Age	−0.005***
	(0.0002)
Height	0.002***
	(0.0001)
Weight	0.001***
	(0.0001)
Constant	−0.270***
	(0.022)
Sex	Yes
Year	Yes
Sport	Yes
N	200,737
Log Likelihood	−61,129.000
Akaike Inf. Crit.	122,413.000
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

7.1 Controlling for Sport

In order to better understand the role age plays, we run regressions on subgroups of sports. We present a table of estimates of Age where the p value is less than 0.05.

Table 6: Impact of higher age on medal probability across sports, statistically significant

	Sport	Age
22	Synchronized Swimming	0.03
19	Figure Skating	0.02
1	Football	0.01
21	Volleyball	0.01
3	Cross Country Skiing	0.01
18	Modern Pentathlon	0.01
24	Nordic Combined	0.01
2	Speed Skating	0.01
9	Rowing	0.01
8	Luge	0.01
7	Alpine Skiing	0.00
15	Diving	0.00
16	Canoeing	0.00
26	Ski Jumping	0.00
17	Tennis	0.00
10	Bobsleigh	0.00
6	Gymnastics	0.00
11	Equestrianism	0.00

	Sport	Age
14	Cycling	0.00
5	Sailing	0.00
4	Athletics	0.00
12	Shooting	0.00
20	Archery	0.00
23	Table Tennis	0.00
13	Taekwondo	-0.01
25	Rugby Sevens	-0.01

We see that Age has a different role to play across sports. While Synchronized Swimming sees the maximum positive impact of age, Rugby Sevenssees the maximum negative impact of age. All sports that have a negative sign, have a negative impact of increasing age, on the likelihood of a medal.

7.2 Controlling for Sport and Year

Table 7: Impact of higher age on medal probability across sports controlling for year, statistically significant

	Sport	Age
19	Figure Skating	0.02
1	Football	0.01
3	Cross Country Skiing	0.01
11	Rowing	0.01
18	Modern Pentathlon	0.01
10	Luge	0.01
2	Speed Skating	0.01
8	Alpine Skiing	0.01
9	Handball	0.00
17	Tennis	0.00
16	Canoeing	0.00
12	Bobsleigh	0.00
13	Equestrianism	0.00
7	Gymnastics	0.00
4	Swimming	0.00
6	Biathlon	0.00
5	Sailing	0.00
14	Shooting	0.00
15	Taekwondo	-0.01

I also control for Years as factors and repeat the analysis. While Figure Skating sees the maximum positive impact of age, Taekwondosees the least impact of age. While controlling for years, we see that there is no sport that has a negative impact on likelihood of medal due to age.

7.3 Controlling for Event and Year

Table 8: Impact of higher age on medal probability across Events controlling for Year, statistically significant

	Event	Age
5	Cross Country Skiing Men's 4 x 10 kilometres Relay	0.02
6	Athletics Women's 4 x 100 metres Relay	0.01
2	Speed Skating Women's 500 metres	0.01
1	Football Men's Football	0.01
3	Cross Country Skiing Men's 10 kilometres	0.01
4	Cross Country Skiing Men's 50 kilometres	0.01
7	Ice Hockey Men's Ice Hockey	0.00

I go one level deeper and look at the same analysis at event level. An event is one rendition of a sport. For example Swimming is a sport but 100m, 200m and 100m backstroke are different events. While sees the maximum positive impact of age, sees the least impact of age. As my data has become granular, there are fewer observations per regression set and the impact of Age has also becomes significant for fewer events.

Table 9: GLM Regression Results on Whether Gold Medal

	Gold Whether Gold Medal
Age	0.0002^{*}
	(0.0001)
Height	0.001^{***}
	(0.0001)
Weight	0.001^{***}
	(0.0001)
Constant	-0.160^{***}
	(0.011)
Gender Controls	Yes
N	206,165
R^2	0.006
Adjusted R^2	0.005
Residual Std. Error	0.220 (df = 206160)
F Statistic	286.000^{***} (df = 4; 206160)
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.