Analysing the development of elite swimmers

Many different events exist in swimming in different formats

- Different length of pool 50m pool vs 25m pool
- Different strokes (butterfly, backstroke, breaststroke, freestyle, individual medley)
- Different distances (50m, 100m, 200m, 400m, etc.)

Questions:

- **How does the progress of elite swimmers vary by gender** and does it line up with what we would expect to see? For example, men hit puberty later than women – are these developmental changes reflected in progression of times?
- **Do race tactics vary by gender** for elite swimmers?
- Individual medley events involve swimmers racing each stroke – butterfly, backstroke, breaststroke, freestyle – in a single race. Are there any **noticeable differences in approaches** used based on gender between elite athletes?

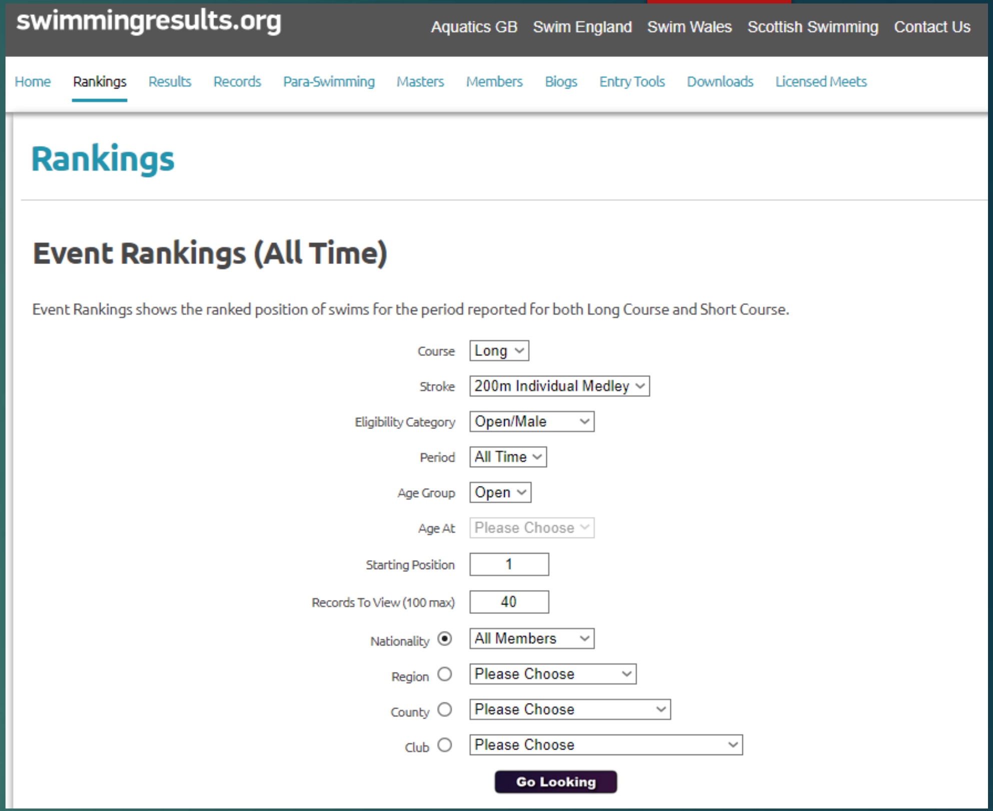
- **These are just a sample of the questions we could ask**. We can make many comparisons based on stroke, age and gender and consider so many different analytics based on progressions on athletes.

Why are such questions important?

- Analysing pb progressions of elite swimmers as they age and comparing them can **help identify trends in the rate at which elite athletes tend to develop**. This can be compared across stroke and distance.
- Analysing race tactics can be done by comparing split times from top perfomers, i.e. getting **each 50m or 100m split from a 200m** swimmer. This can be scaled by finding them as percentages of the total time so that comparisons can be made across stroke and gender.
- **By understanding common trends and groupings of athletes** and how they develop, it is possible that such analytics and models can help coaches, swimmers and parents alike devise plans and tactics to **help drive improvements and progress**, as well as to understand why swimmers may typically not be performing as well as ideal.

Datasets

- British Swimming Event Rankings – a very comprehensive database which is constantly being updated with meet results. It contains swims and swimmer profiles, with swims being attached to profiles. From profiles and cross-referencing, ages can be calculated, as well as dates put to swims, such that models can be developed.
- Plenty of records are accessible, dating back as far as 1997, although many meets are unlicensed and don't get put onto rankings – this issue is more noticeable in the earlier years (pre-2010, for example)
- Using data from too far back may have too many gaps in PB progressions of swimmers due to fewer meets being uploaded to rankings
- The data is all online, publicly available and accessible

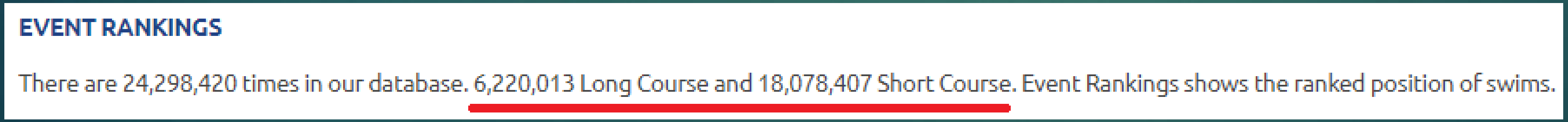


Rankings

Event Rankings (All Time)

Event Rankings shows the ranked position of swims for the period reported for both Long Course and Short Course.

Course: Long
Stroke: 200m Individual Medley
Eligibility Category: Open/Male
Period: All Time
Age Group: Open
Age At: Please Choose
Starting Position: 1
Records To View (100 max): 40
Nationality ● All Members
Region ○ Please Choose
County ○ Please Choose
Club ○ Please Choose



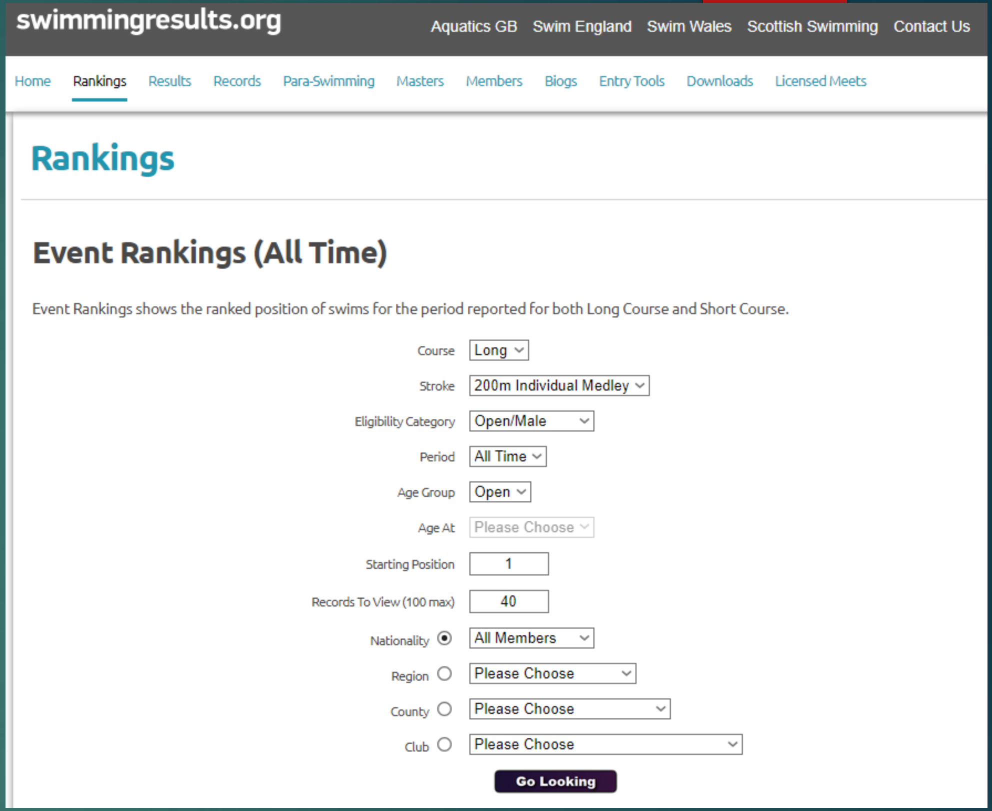
EVENT RANKINGS

There are 24,298,420 times in our database. 6,220,013 Long Course and 18,078,407 Short Course. Event Rankings shows the ranked position of swims.

https://www.swimmingresults.org/eventrankings/

Datasets

- As we can see, there is a great level of customisability for querying the database. We can query based on:
- the length of the pool used for a race (Long course vs Short course)
- Stroke and event
- Gender aka eligibility category
- The year to search for swims
- Age – open is all ages; age at specifies whether the swims should be found based on a swimmer's age on the day of the swim or their age at the end of the year – this latter option is more relevant as **almost all major competitions and team selections in the UK are based on age at end of year**
- We can even localise by region, county and club
- Nationality – this aspect is important for a few reasons. The main reason is that a reasonable portion of swimmers who could be considered elite (national finalists) are from abroad and board at private schools or train at university. This means records of their pb progressions aren't always available in their history on the site, so attempting to use swimmers in calculations and modeling could introduce anomalies and outliers.



Rankings

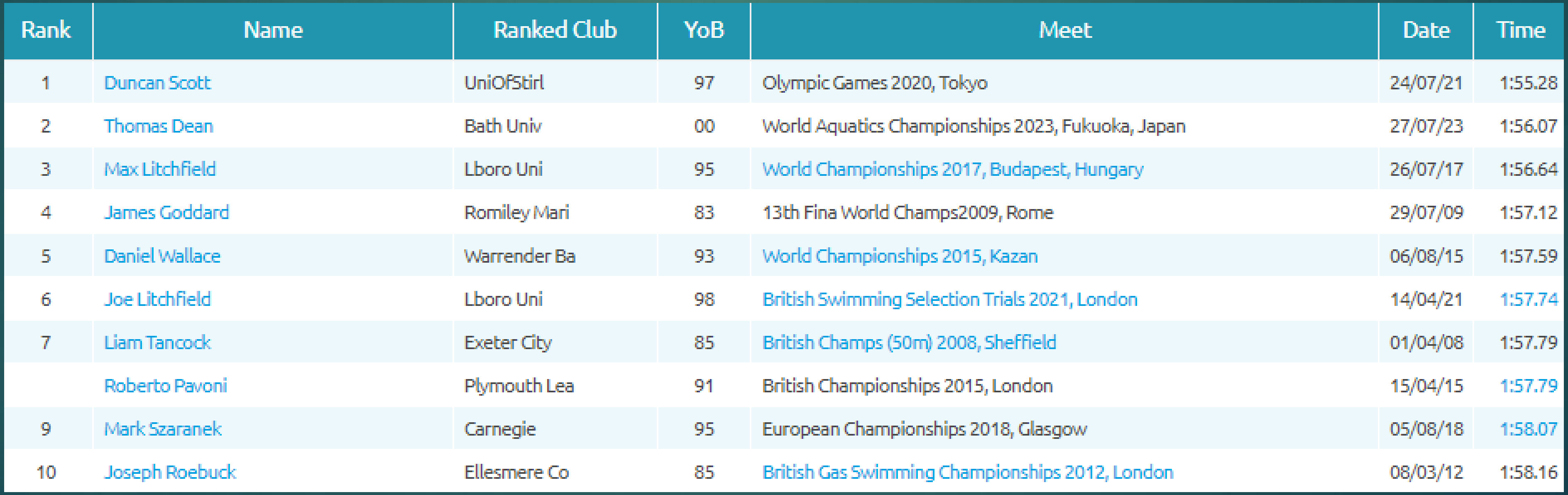
Event Rankings (All Time)

Event Rankings shows the ranked position of swims for the period reported for both Long Course and Short Course.

Course: Long
Stroke: 200m Individual Medley
Eligibility Category: Open/Male
Period: All Time
Age Group: Open
Age At: Please Choose
Starting Position: 1
Records To View (100 max): 40
Nationality ● All Members
Region ○ Please Choose
County ○ Please Choose
Club ○ Please Choose

Datasets

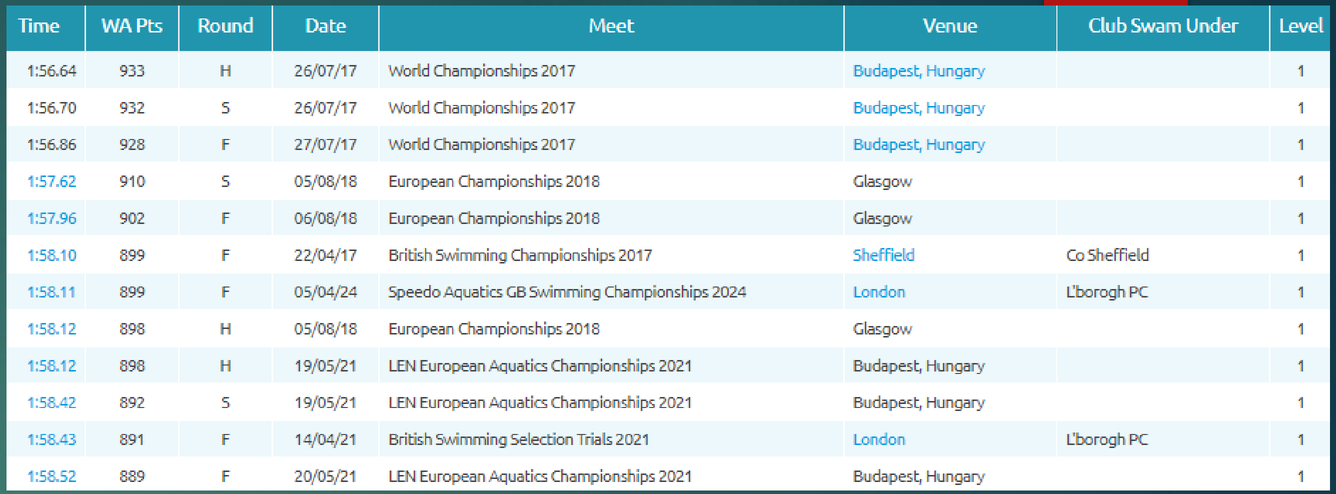
- Here is an example of a query output based on the settings from the last slide.
- Notice how each name is a hyperlink – within the HTML, these contain their ID numbers which can be used to get tables of their best times



Rank	Name	Ranked Club	YoB	Meet	Date	Time
1	Duncan Scott	UniOfStirl	97	Olympic Games 2020, Tokyo	24/07/21	1:55.28
2	Thomas Dean	Bath Univ	00	World Aquatics Championships 2023, Fukuoka, Japan	27/07/23	1:56.07
3	Max Litchfield	Lboro Uni	95	World Championships 2017, Budapest, Hungary	26/07/17	1:56.64
4	James Goddard	Romiley Mari	83	13th Fina World Champs2009, Rome	29/07/09	1:57.12
5	Daniel Wallace	Warrender Ba	93	World Championships 2015, Kazan	06/08/15	1:57.59
6	Joe Litchfield	Lboro Uni	98	British Swimming Selection Trials 2021, London	14/04/21	1:57.74
7	Liam Tancock	Exeter City	85	British Champs (50m) 2008, Sheffield	01/04/08	1:57.79
	Roberto Pavoni	Plymouth Lea	91	British Championships 2015, London	15/04/15	1:57.79
9	Mark Szaranek	Carnegie	95	European Championships 2018, Glasgow	05/08/18	1:58.07
10	Joseph Roebuck	Ellesmere Co	85	British Gas Swimming Championships 2012, London	08/03/12	1:58.16

Datasets

- Swims can be sorted in time order or date order – this can be useful depending on the data we want to collect, i.e. for pb progressions, we probably want in date order so that we can filter out times which aren't PBs and have a ready sequence. For comparing race tactics, selecting the top n times may be more effective.
- As you can see, some swim times are hyperlinks – this is where splits are available and can be scraped.



Time	WA Pts	Round	Date	Meet	Venue	Club Swam Under	Level
1:56.64	933	H	26/07/17	World Championships 2017	Budapest, Hungary		1
1:56.70	932	S	26/07/17	World Championships 2017	Budapest, Hungary		1
1:56.86	928	F	27/07/17	World Championships 2017	Budapest, Hungary		1
1:57.62	910	S	05/08/18	European Championships 2018	Glasgow		1
1:57.96	902	F	06/08/18	European Championships 2018	Glasgow		1
1:58.10	899	F	22/04/17	British Swimming Championships 2017	Sheffield	Co Sheffield	1
1:58.11	899	F	05/04/24	Speedo Aquatics GB Swimming Championships 2024	London	L'borogh PC	1
1:58.12	898	H	05/08/18	European Championships 2018	Glasgow		1
1:58.12	898	H	19/05/21	LEN European Aquatics Championships 2021	Budapest, Hungary		1
1:58.42	892	S	19/05/21	LEN European Aquatics Championships 2021	Budapest, Hungary		1
1:58.43	891	F	14/04/21	British Swimming Selection Trials 2021	London	L'borogh PC	1
1:58.52	889	F	20/05/21	LEN European Aquatics Championships 2021	Budapest, Hungary		1

For scraping the data, I used BeautifulSoup4 in Python which scrapes websites as a tree of HTML DOM elements. I then used pipelines of functions incorporating maps and filters to sort rows and filter out anomalies, such as if swims don't have splits for example.

I've uploaded the base data that I scraped and used for modelling to GitHub:

https://github.com/SamArrows/UK-Swimming-Rankings-Data-Analysis/tree/main/CSV/Files%20for%20Learning%20from%20Data

PB Progressions of Elite Swimmers

- Collecting this data involves getting dates when a swim was performed, the time and then calculating how much of a percentage improvement on the last pb it was.
- The earliest age a swimmer can possibly qualify for a national-level event in the UK is aged 12 at the end of the year – this is a good baseline to start finding PBs from.
- As swimmers' birthdays aren't accessible, a uniform way to evaluate swimmers' progressions over time would be to start at the earliest point of the year in which they turn 12
- Passage of time in days is an independent variable so would go on the x-axis, whilst pb percentage change can be plotted on the y-axis as this variable is dependent on how much more developed the swimmer is than their last swim

Polynomial Regression

- As we can see, a good technique to use for modelling PB changes over time would be to develop a regression model as both variables are continuous. Using **Bayes Information Criterion**, we can determine its degree.
- We could develop a regression model for, say, the top 40 athletes in a given event in a given year for males and females and compare them.
- Such a model can give us a general idea of how male performances vary with female performances as they age for a given event. Do men keep seeing large percentage changes in their swims compared to women as they age?

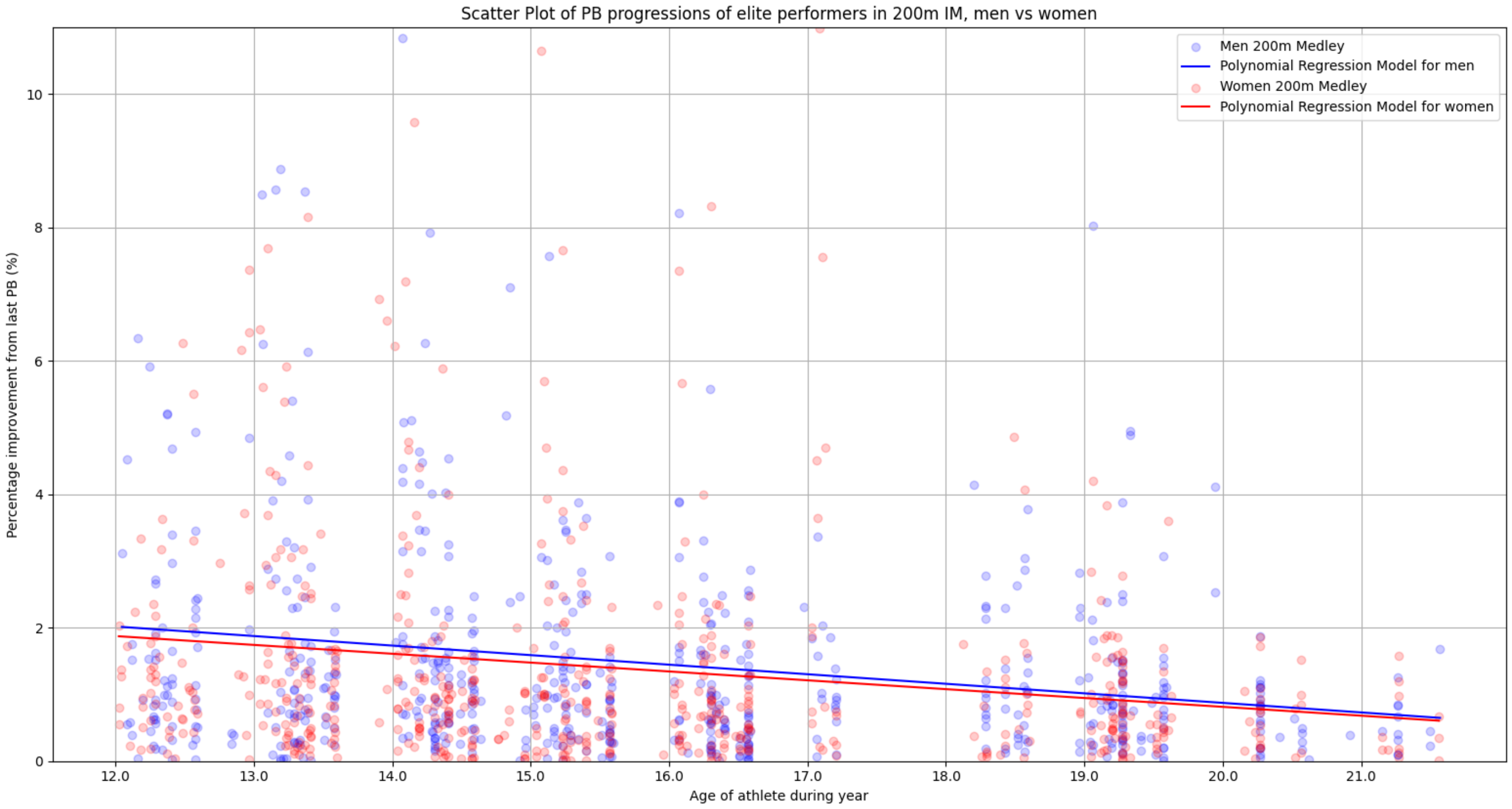
$$\mathrm{BIC} = k\ln(n) - 2\ln(\widehat{L}).$$

- In the formula: L is the maximum likelihood estimate of the likelihoodfunction, k is the number of parameters in the model, and n is the number of data points

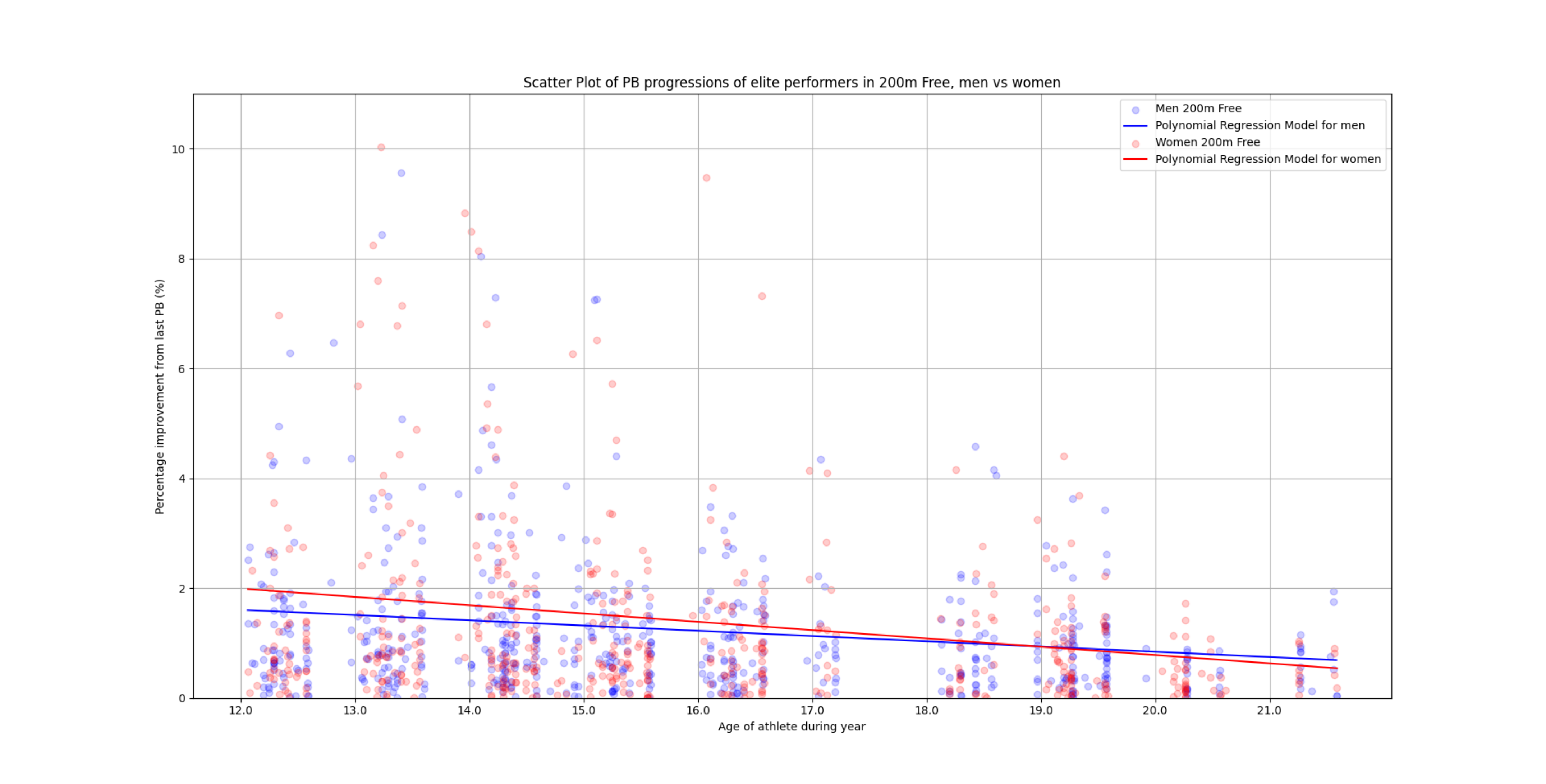
Data to apply regression to:

- We need to ensure there are plenty of swims as the swimmer has gone through puberty, hence if we are starting at age 12, we ideally need to collect records not much earlier than 2010 as records can be missing when we search for earlier years
- We want to see recent trends and consider the latest trends with respect to the latest technique developments, training programmes, diets, etc.
- Let us select the top 40 swimmers (male and female) from 2022 as many of these will have been in their teens through the 2010s – plenty of data as well as being recent.
- Consider 200m Freestyle and 200m Medley

Linear Regression of 200 IM



Linear Regression of 200m Free



Interesting things to note:

- There is a gap between each year in the latter months (Sept-Dec) which intuitively corresponds to the short course season where athletes don't race long course
- This is more pronounced in the 17-18 age group, which likely correlates to adapting to university life or studying A-Levels/exams, hence why we see another cluster after this stage
- Both models start with 2% change in PB at age 12 for both genders which tapers towards 0.6-0.7% around age 21, however, interestingly, women see bigger percentage changes up to around 18 in freestyle than men according to the model, with this dropping below the mens rate after this age, whereas in medley, men consistently have a higher percentage change in PB compared to women up to 21 – clear difference in development based on strokes
- There are a lot of outliers, which regression models can be sensitive to. Even though I removed those above 10%, there are a lot in the early 12-14 year old stages between 5-10% - likely correlates to big changes in lifestyle and developing

Clustering by race tactics

- Suppose we want to consider how top athletes tend to split their 200m races: do they emphasise front-end speed with a proportionally far faster first 100m or do they take the first 100m easier to push harder in the second 100m? Taking this into account, how far off their PB were they using such tactics?

- For simplicity, let us focus just on the 200m Medley. To get data, we need to ensure we are evaluating elite athletes in peak form. By considering all-time best performers, we can get a tighter group of elite athletes. To ensure plenty of data, let us select the top 40 and choose swims which are within 2% of their PB time, to ensure swims are amongst their best performances
- If we plot 100m time as percentage of total swim time on the x-axis, we can plot how far off PB they were as a percentage on the y-axis – this will be capped at 2% obviously

Techniques to use:

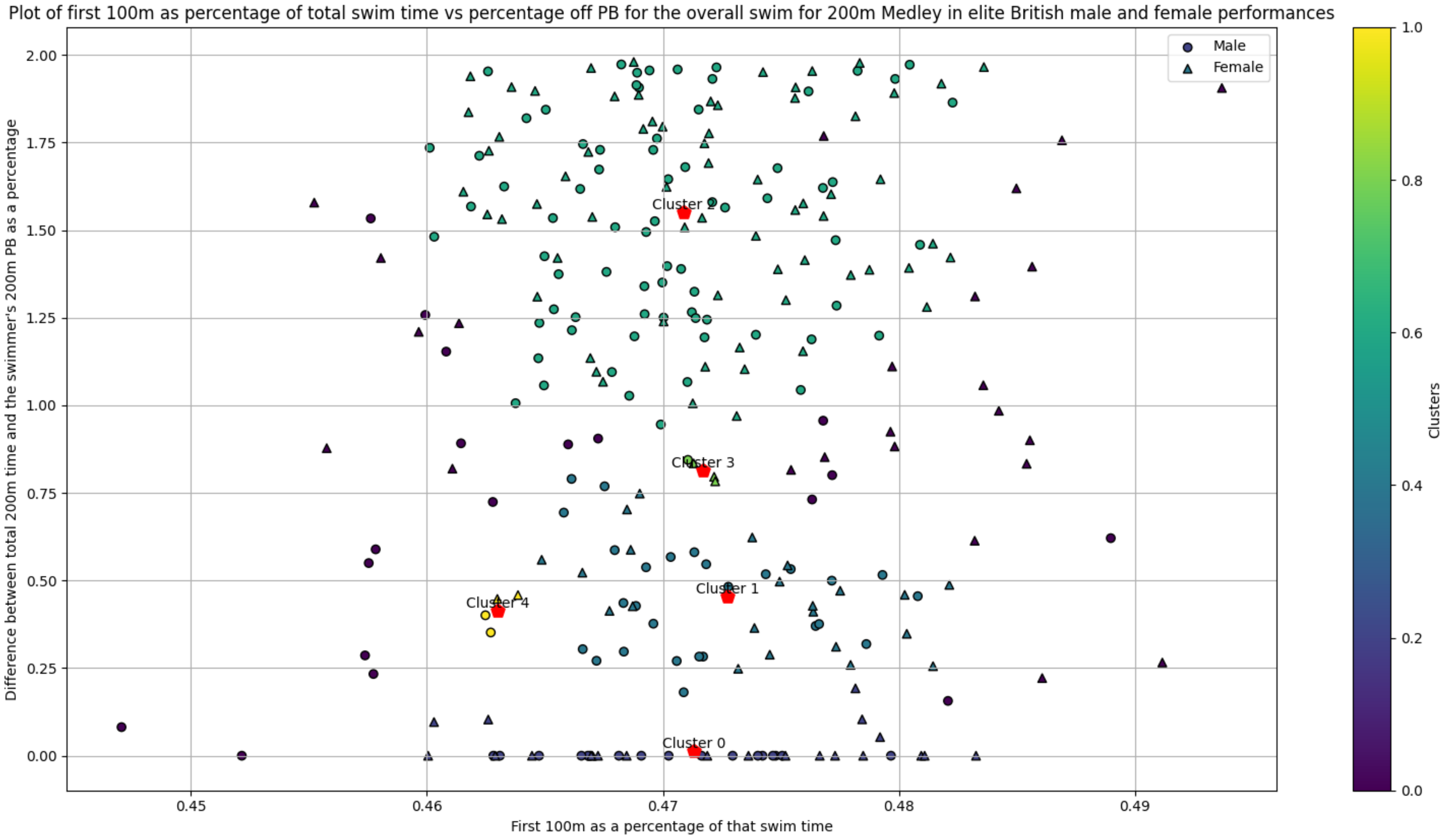
- I tried several techniques but noticed that the spread of men and women was even across the whole space, implying that tactics are similar but also that determining clusters would be difficult to do
- To determine clusters, we need to evaluate them on density rather than just distance metrics
- We also need to identify outliers and tactics which didn’t seem as common
- K-nearest-neighbour is not inherently for clustering and focuses on classification, as well as being supervised; K-means is unsupervised but requires the number of clusters and the starting means of the clusters to be predetermined, which we can’t reliably specify
- Both are skewed by outliers and due to needing some label or predetermined info on the clusters beforehand, are unsuitable compared to DBSCAN which is unsupervised, robust to noise, and calculates clusters by itself using densities of points

DBSCAN

- DBSCAN requires data to be scaled due to using radius and minimum points within radii to determine its clusters
- Our scale on the x-axis is within 0.4-0.5 whereas y-axis is going to be 0.0-2.0, hence we need to rescale for applying the algorithm but can plot the original scaled points once they have been assigned clusters
- DBSCAN can yield clusters of varying shapes so finding mean averages of clusters may not be very representative of them, however if our clusters turn out to be uniformly shaped then it could be informative to plot mean points

DBSCAN

Used MinMax scaling, with eps=0.06 and min_samples=5

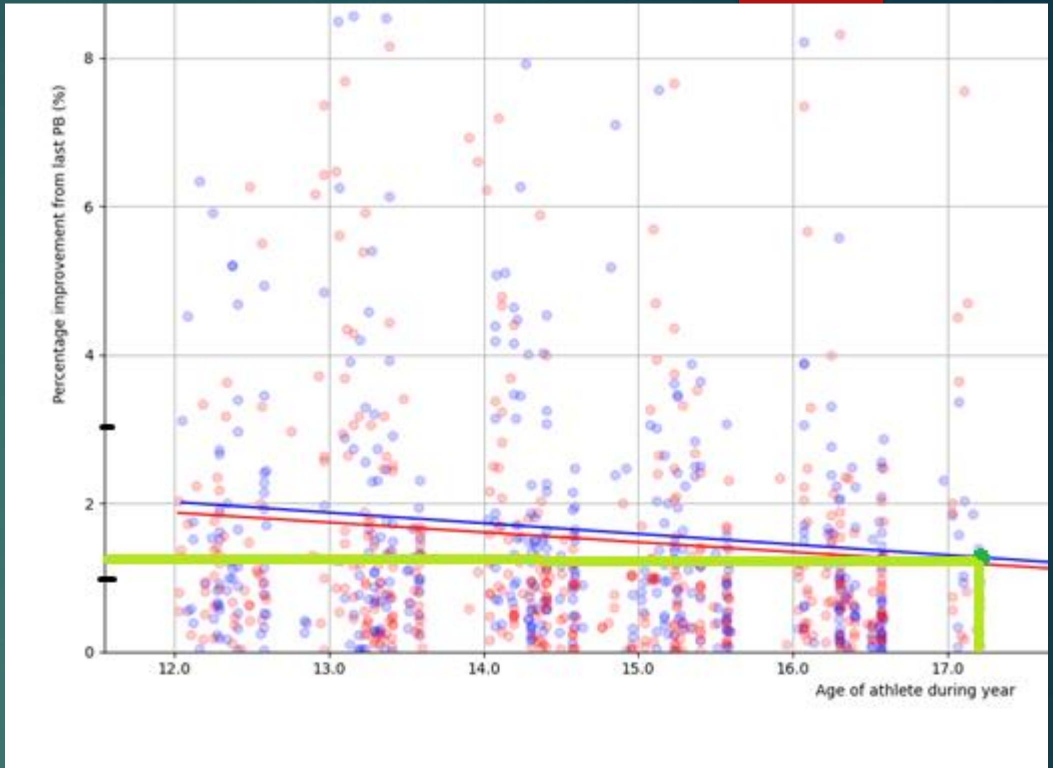


Interesting things to note:

- The main thing to note with the clusters was that they mostly ranged between 46-48% for the first 100m when excluding outliers, with the mean points of the fastest cluster (close to 0% off PB) being just above 47%.
- As the range for the x-axis is consistent for most of the clusters, it seems that this range of 46-48% is the sweet spot for how to split a 200m Medley – remember that we selected swims within 2% as these are all strong performances, so the cluster closest to 0% may be the 'best' cluster to aim to be in but the tactics in the whole plot are all relevant. The fact they are seemingly quite similar is promising.
- As the clusters are uniform shapes – roughly, we could approximate them as ovals/ellipses – we can plot mean points for them

Conclusion

- So how is this data important???
- Let us work through devising a race tactic for a swimmer. Suppose we have a 16-year-old boy who races 200m Medley in 2:10.00 = 130 seconds. What would be a realistic target time to aim for next season when he is 17?
- Using the regression model, the rate of change in PB can be approximated at around 1.5%
- 130 * 0.015 = 1.95; 130 - 1.95 = 128.05 (2:08.05)
- The clustering model shows that when the best 200m medley performers are on PB pace, their first 100m makes up between 46-48% of the swim, with the mean average being around 47.1% of the swim:
- 0.471 * 128.05 = 60.31 (1:00.31)
- Hence, using multiple models derived from the Event Rankings database, we can predict seasonal targets and progressions for developing athletes.
- Further modelling could be done to refine the predictions and provide even more insights, such as modeling top performers in specific age groups and comparing split percentages with elite open-age performers.


Percentage improvement from last PB (%)
Age of athlete during year