Analyzing Current world Leaders in Swimming to World Record Holders

Eric Kim

2025-06-30

Table of contents

Introduction

Swimming is very popular sport where althetes will race against each other in one of four strokes.It has become so recognized that an international federation called FINA was created and started to regulate rules. Rules like technique, suits you can wear, drugs you can and cant use. Since the early 1900's, these rules have slowly changed over time, and so has the records for each event.

In this project, I want to find more about 3 main questions:

1. How far away are the top times from the current world record?

2. What countries did the best during the Olympics throughout 1912-2020?
3. What are the differences between 2 Olympic Games times?

These questions will help find how hard a record is to break as well as see how much the average time has gotten over the years. Some records have lasted for decades, while others have recently been broken. By analyzing the times and when they occurred we can see just how close swimmers are to really breaking the record.

Data Provenance

For this project, I used 3 data sets to explore how well swimmers perform during the Olympics and how they compare to a World record Holder. The data sets were from a reliable source, mainly Kaggle. The data used was directly from the Olympics website. Each cases is determined by the event and the time swam.

Primary Dataset

- Source: Kaggle
- Description: The primary dataset consists of detailed swims that occured during any Olympics from 1912-2020. The content most likely used are the host city of olympics, the year, distance and stroke of event, gender, names, and results. It has 4,359 rows and 10 cloumns to represent each performance during the olympics.

Secondary Dataset

- Source: Wikipedia
- Description: This secondary data set provides us with all possible world record times for men and women, as well as, long course and short course. I will be focusing on the short course tables. Like the primary data set, this data set includes the event, time, name, and location. it contains 24 rows and 8 columns.
- Source: Kaggle
- Description: This secondary data set provides the best 200 times in each swim event. It contains event name, time, athlete name, gender, rank order,etc. It has 5200 rows and 14 columns

```
Attaching package: 'dplyr'
```

```
The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union


Attaching package: 'rvest'

The following object is masked from 'package:readr':

    guess_encoding

Rows: 4359 Columns: 10
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (7): Location, Distance (in meters), Stroke, Gender, Team, Athlete, Results
dbl (3): Year, Relay?, Rank

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Event	Time	Name	Nationality	Date	Meet	Location	Ref
50m freestyle	19.90sf	Jordan Crooks	Cayman Islands	14 December 2024	World Championships	Budapest, Hungary	[97][98]
100m freestyle	44.84	Kyle Chalmers	Australia	29 October 2021	World Cup	Kazan, Russia	[99][100][101]
200m freestyle	1:38.61	Luke Hobson	United States	15 December 2024	World Championships	Budapest, Hungary	[102][103]

Event	Time	Name	Nationality	Date	Meet	Location	Ref
400m freestyle	3:32.25	Yannick Agnel	France	15 November 2012	French Nationals	Angers, France	[104][105]
800m freestyle	7:20.46	Daniel Wiffen	Ireland	10 December 2023	European Championships	Otopeni, Romania	[106][107]
1500m freestyle	14:06.88	Florian Wellbrock	Germany	21 December 2021	World Championships	Abu Dhabi, United Arab Emirates	[108][109]
50m backstroke	22.11	Kliment Kolesnikov	Russia	23 November 2022	Solidarity Games	Kazan, Russia	[110][111][112]
100m backstroke	48.33	Coleman Stewart	United States	29 August 2021	International Swimming League	Naples, Italy	[113][114]
200m backstroke	1:45.63	Mitch Larkin	Australia	27 November 2015	Australian Championships	Sydney, Australia	[115][116]
50m breaststroke	24.95	Emre Sakçı	Turkey	27 December 2021	Turkish Championships	Gaziantep, Turkey	[117][118]
100m breaststroke	55.28	Ilya Shymanovich	Belarus	26 November 2021	International Swimming League	Eindhoven, Netherlands	[119]
200m breaststroke	2:00.16	Kirill Prigoda	Russia	13 December 2018	World Championships	Hangzhou, China	[120][121]

Event	Time	Name	Nationality	Date	Meet	Location	Ref
50m butterfly	21.32	Noè Ponti	Switzerland	11 December 2024	World Championships	Budapest, Hungary	[122][123]
100m butterfly	47.71	Noè Ponti	Switzerland	14 December 2024	World Championships	Budapest, Hungary	[124][125]
200m butterfly	1:46.85	Tomoru Honda	Japan	22 October 2022	Japanese Championships	Tokyo, Japan	[126][127]
100m individual medley	49.28	Caeleb Dressel	United States	22 November 2020	International Swimming League	Budapest, Hungary	[128][129]
200m individual medley	1:48.88	Léon Marchand	France	1 November 2024	World Cup	Singapore, Singapore	[130][131]
400m individual medley	3:54.81	Daiya Seto	Japan	20 December 2019	International Swimming League	Las Vegas, United States	[132][133]
4 × 50 m freestyle relay	1:21.80	Caeleb Dressel (20.43)Ryan Held (20.25)Jack Conger (20.59)Michael Chadwick (20.53)	United States	14 December 2018	World Championships	Hangzhou, China	[134]
4 × 50 m freestyle relay	1:20.77 WB ss	[135] Bernard (20.64)Fabien Gilot (20.33)Amaury Leveaux (19.93)Frédérick Bousquet (19.87)	France	14 December 2008	European Championships	Rijeka, Croatia	[136]

Event	Time	Name	Nationality	Date	Meet	Location	Ref
4 × 100 m freestyle relay	3:01.66	Jack Alexy (45.05)Luke Hobson (45.18)Kieran Smith (46.01)Chris Guiliano (45.42)	United States	10 December 2024	World Championships	Budapest, Hungary	[137][138]
4 × 200 m freestyle relay	6:40.51	Luke Hobson (1:38.91)Carson Foster (1:40.77)Shaine Casas (1:40.34)Kieran Smith (1:40.49)	United States	13 December 2024	World Championships	Budapest, Hungary	[139][140]
4 × 50 m medley relay	1:29.72	Lorenzo Mora (22.65)Nicolò Martinenghi (24.95)Matteo Rivolta (21.60)Leonardo Deplano (20.52)	Italy	17 December 2022	World Championships	Melbourne, Australia	[141]
4 × 100 m medley relay	3:18.68	Miron Lifintsev (49.31)Kirill Prigoda (55.15)Andrei Minakov (48.80)Egor Kornev (45.42)	Russia	15 December 2024	World Championships	Budapest, Hungary	[142][143]

```
Rows: 5200 Columns: 14
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (12): Event Name, Swim time, Swim date, Event description, Team Code, Te...
dbl  (2): index, Rank_Order

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Exploratory Data Analysis (EDA)

For my Exploratory Data Analysis(EDA), I wanted to specalize my search from the date to specifically male Butterfly events, in order to answer my research questions. I started by creating a scatter plot of the top 10 swims in a given fly event, the lowest point being the fastest time. A scatterplot allows us to see where some swims line up with others, as well as, seeing if a certain swimmer is setting multiple top times. Along side, I would like to take the top 3 swimmers to create a boxplot. We can use this to see if the top races were a one off event or if they consistently hit a certain mark.

Scatter Plots

This scatter plot(Figure 1) will show the relationship between age of athlete to the times they achieved.

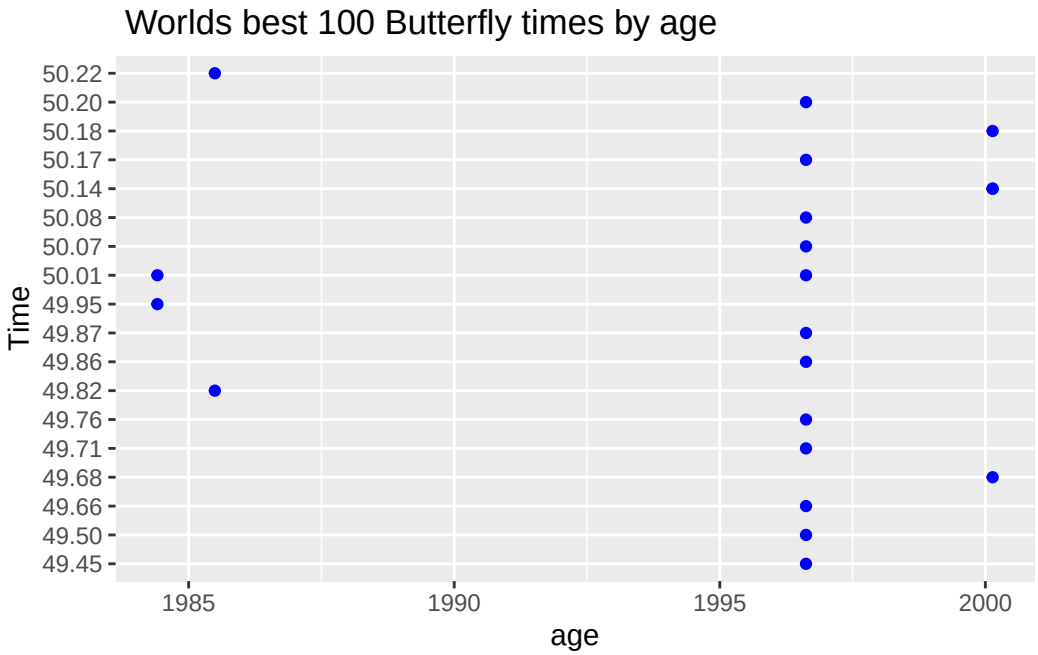

Figure 1: Scatter Plot of ages that produced the fastest times in 100 Butterfly.

Thoughts: Looking at the scatterplot, we can clearly see one age stood out way more than the rest. The others typically stayed in the top half of the plot(slower times) with a couple one of races. Since the plot is only sampling from the top 20 times in this event, we can see that there was a long period of time where the lower of the 1986 points stayed as World Leader. This allows us to see just how long it took for a new World Record Swimmer to emerge and just how unlikely it is to be broken any time soon.

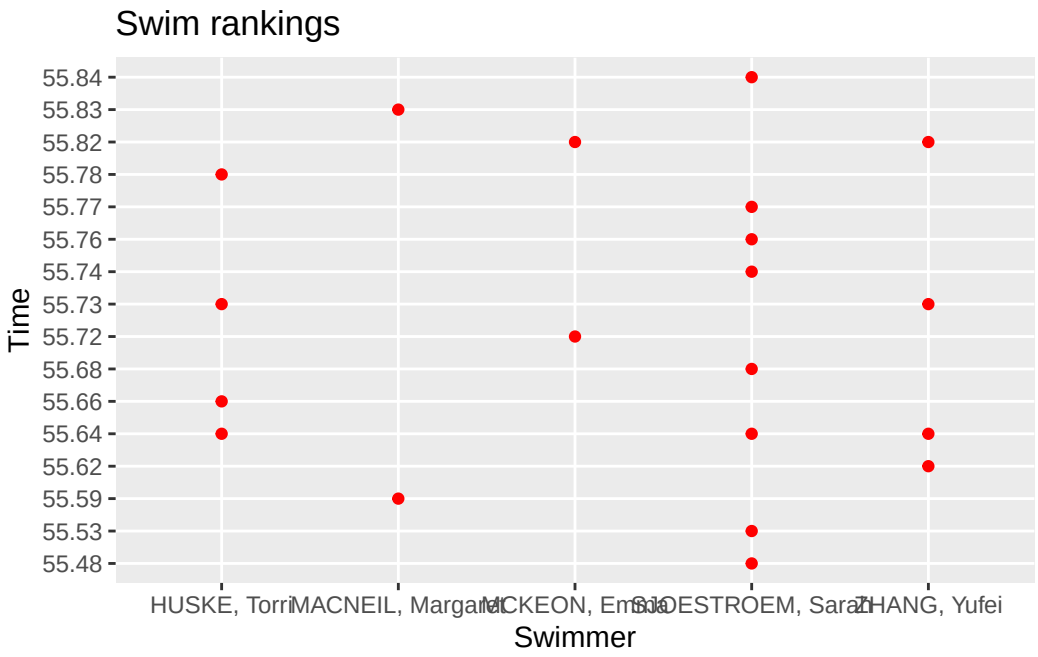

Figure 2: Scatter Plot of the top 20 Women's 100 Butterly times by name

Thoughts: This gives a great insight on which events are being dominated and which is an even competition. The plot shows that the top 20 times are all from 5 swimmers, with one(the WR holder) being having an overwhelming amount of top times. We can see how some swimmers are very consistent, while others very in time. By analyzing further, we can see if other factors come into play or if it was purely just a good race. Some factors include: Location, time of year, time in between top swims.

Box Plots

This box plot will show us whether Men or Women have a larger gap between top 5 swims for 100 Butterfly.

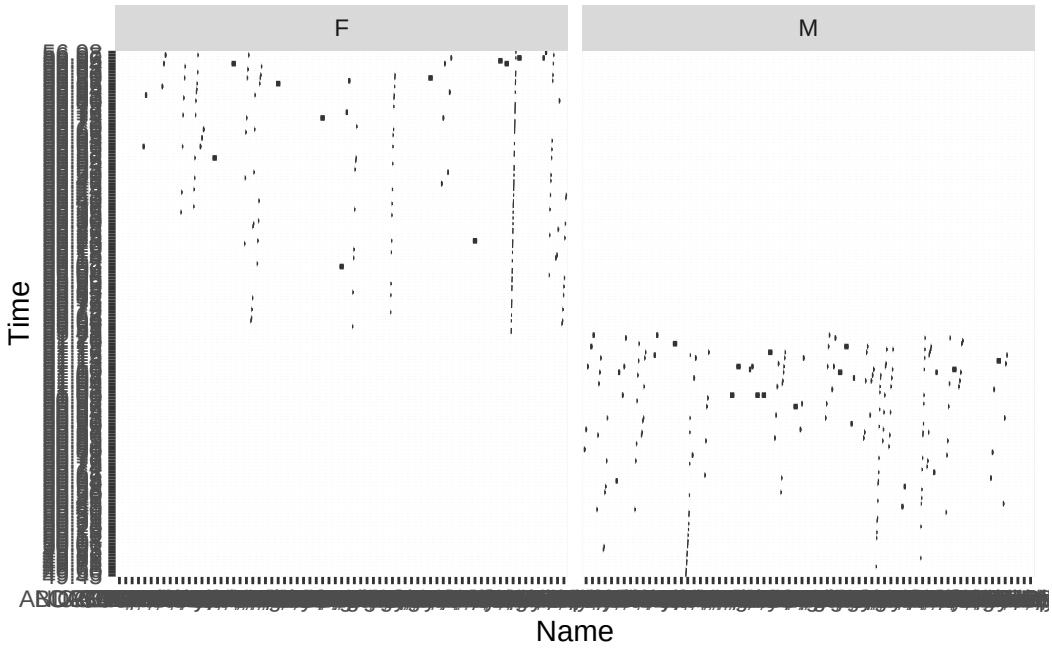

Figure 3: Box Plot showing the gap between World leaders and their top competiters

Thoughts: This lets us see how the top 200 times for 100 butterfly stack up against each other. The bolder the boxplot, the more frequent the time was hit. We are able to see how much some of the fast swimmers vary from race to race. Thhere are some swimmers who stay toward the slower part of the range, while others have a whole array that spans from one of the slowest of the 200 swims and one of the fastest. One good race doesn't always determine everything. There may be a outlier swim that isn't reproducable.

Conclusion

We are way further than we think we are from break many of the current world Records. Many of the possible candidates are either getting past their prime or retired. We need to wait for the need huge wave of talent to take over the swimming scene. As seen in the plots, many of the top times are controlled by a select few. In order for things to get more competitive, the floor for performance must be raised, for both Male and Female.

Sources

"List of World Records in Swimming." Wikipedia, Wikimedia Foundation, 20 June 2025, en.wikipedia.org/wiki/List_of_world_records_in_swimming.

"Olympic Swimming History (1912 to 2020)." Kaggle, 18 Apr. 2023, www.kaggle.com/datasets/datasciencedonut swimming-1912-to-2020/data.

"Swimming: Top in the World in Each Category." Kaggle, 27 Nov. 2022, www.kaggle.com/datasets/thedevastato top-200-world-times-in-each-category.

Code Appendix

```
#Tidyverse Style Guide
library(dplyr)
library(knitr)
library(ggplot2)
library(readr)
library(rvest)

# Load the primary data set
primary_data <- read_csv("/Users/erickim/Desktop/Olympic_Swimming_Results_1912to2020 (1).csv"


# Load the 1st secondary data set
url1 <- "https://en.wikipedia.org/wiki/List_of_world_records_in_swimming"
read <- read_html(url1)

ListOfTables <- html_table(read,fill = TRUE)
#short course men's world records
recordTable <- ListOfTables[[4]]
kable(recordTable)
# Load the 2nd secondary data set
Swimming_database_2 <- read_csv("/Users/erickim/Desktop/Swimming database 2.csv")


# Rename Columns
primary_data <- primary_data %>%
  rename(Name = "Athlete")
Swimming_database_2 <-Swimming_database_2%>%
  rename(
```

```
    Name = "Athlete Full Name",
    Time = "Swim time",
    Event = "Event description",
    Rank = "Rank_Order",
    date = "Swim date",
    Event_name = "Event Name",
    Team_name = "Team Name",
    age = "Athlete birth date",
    gender = "Gender"
    )

# Merge Data sets
merged_data <- Swimming_database_2 %>%
  left_join(
    primary_data,
    by = "Name"
  )

# Scatter Plot1
Swimming_database_2$age<- as.Date(Swimming_database_2$age, format = "%m-%d-%y")
new_age <- format(Swimming_database_2$age, "%y-%m-%d")
byAge <- Swimming_database_2%>%
  filter( Event %in% c( "Men 100 Butterfly LCM Male"))%>%
  arrange(Time)%>%
  filter(row_number()<=20)

ggplot(byAge, aes(x = age, y = Time))+
  geom_point( color = "blue")+
  labs(
    title = " Worlds best 100 Butterfly times by age"
  )

# Scatter Plot2
byName <- Swimming_database_2%>%
  filter( Event %in% c( "Women 100 Butterfly LCM Female"))%>%
  filter(row_number()<=20)

ggplot(byName, aes(x =Name, y = Time)) +
  geom_point(color = "red") +
  labs(title = "Swim rankings", x = "Swimmer", y = "Time")

# Boxplot
```

```r
boxplot1 <- Swimming_database_2%>%
  filter( Event %in% c( "Men 100 Butterfly LCM Male","Women 100 Butterfly LCM Female"))

ggplot(boxplot1, aes(x =Name , y = Time)) +
  geom_boxplot() +
  facet_wrap(~gender)+
  labs(
    x = "Name",
    y = "Time"
  )

library(dplyr)
library(knitr)
library(ggplot2)
library(readr)
library(rvest)

# Load the primary data set
primary_data <- read_csv("/Users/erickim/Desktop/Olympic_Swimming_Results_1912to2020 (1).csv"


# Load the 1st secondary data set
url1 <- "https://en.wikipedia.org/wiki/List_of_world_records_in_swimming"
read <- read_html(url1)

ListOfTables <- html_table(read,fill = TRUE)
#short course men's world records
recordTable <- ListOfTables[[4]]
kable(recordTable)
# Load the 2nd secondary data set
Swimming_database_2 <- read_csv("/Users/erickim/Desktop/Swimming database 2.csv")


# Rename Columns
primary_data <- primary_data %>%
  rename(Name = "Athlete")
Swimming_database_2 <-Swimming_database_2%>%
  rename(
    Name = "Athlete Full Name",
    Time = "Swim time",
    Event = "Event description",
    Rank = "Rank_Order",
```

```
    date = "Swim date",
    Event_name = "Event Name",
    Team_name = "Team Name",
    age = "Athlete birth date",
    gender = "Gender"
    )

# Merge Data sets
merged_data <- Swimming_database_2 %>%
  left_join(
    primary_data,
    by = "Name"
  )

# Scatter Plot1

Swimming_database_2$age<- as.Date(Swimming_database_2$age, format = "%m-%d-%y")
new_age <- format(Swimming_database_2$age, "%y-%m-%d")
byAge <- Swimming_database_2%>%
  filter( Event %in% c( "Men 100 Butterfly LCM Male"))%>%
  arrange(Time)%>%
  filter(row_number()<=20)

ggplot(byAge, aes(x = age, y = Time))+
  geom_point( color = "blue")+
  labs(
    title = " Worlds best 100 Butterfly times by age"
  )

# Scatter Plot2
byName <- Swimming_database_2%>%
  filter( Event %in% c( "Women 100 Butterfly LCM Female"))%>%
  filter(row_number()<=20)

ggplot(byName, aes(x =Name, y = Time)) +
  geom_point(color = "red") +
  labs(title = "Swim rankings", x = "Swimmer", y = "Time")

# Boxplot
boxplot1 <- Swimming_database_2%>%
  filter( Event %in% c( "Men 100 Butterfly LCM Male","Women 100 Butterfly LCM Female"))
```

```r
ggplot(boxplot1, aes(x =Name , y = Time)) +
  geom_boxplot() +
  facet_wrap(~gender)+
  labs(
    x = "Name",
    y = "Time"
  )
```