Cycosports Race Analysis

Analysis of a publicly available race timings dataset.

Eric Lim

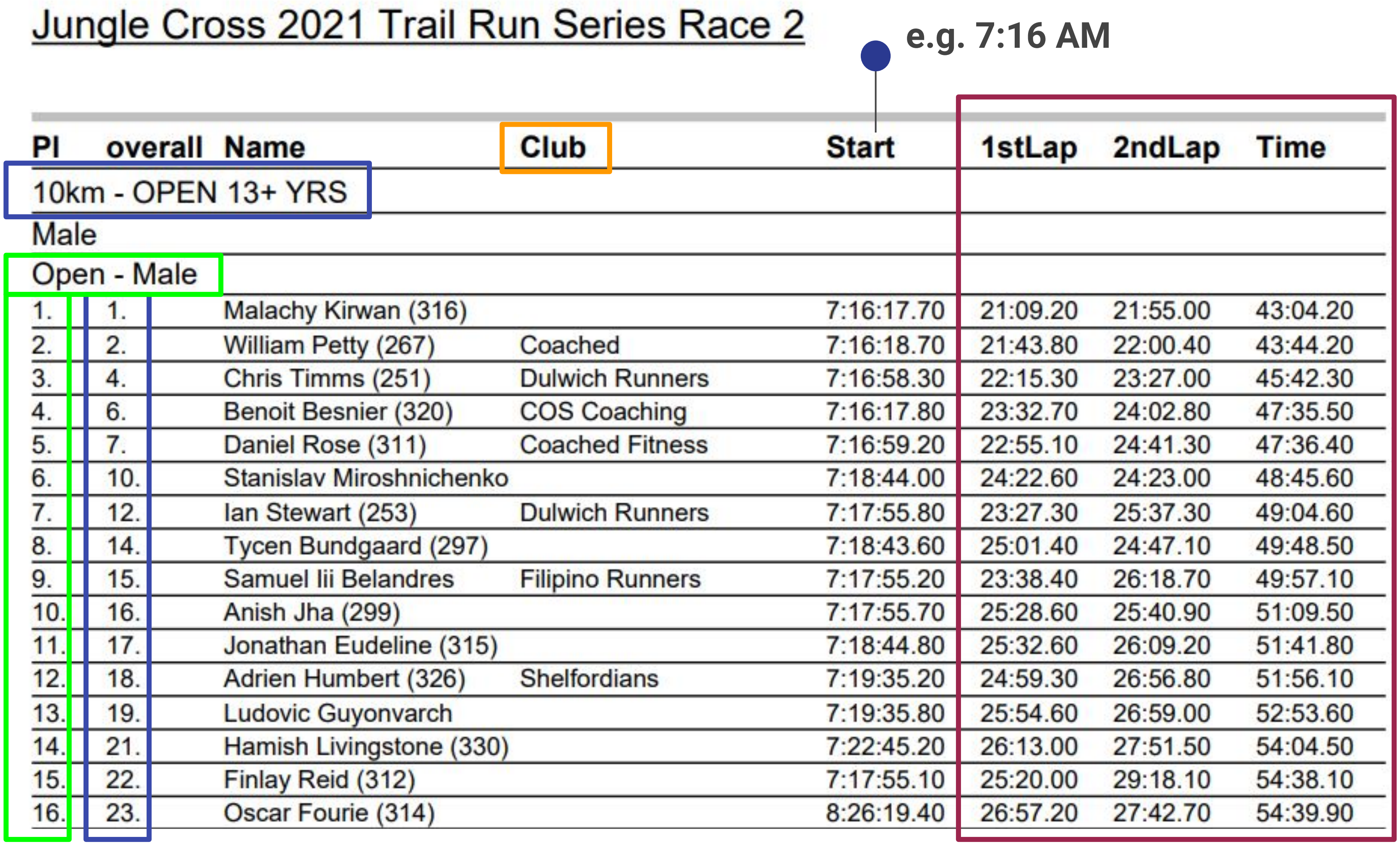
Singapore’s leading organiser of off-road cycling and running events

Singapore's Most Inspiring Trail Runners – Part 2

BY SAMANTHA KHOO ON JAN 21, 2020

Embrace the trail so that it becomes easier to overcome.

runsociety.com



Jungle Cross 2021 Trail Run Series Race 2

e.g. 7:16 AM

Pl	overall	Name	Club	Start	1stLap	2ndLap	Time
10km - OPEN 13+ YRS							
Male							
Open - Male							
1.	1.	Malachy Kirwan (316)		7:16:17.70	21:09.20	21:55.00	43:04.20
2.	2.	William Petty (267)	Coached	7:16:18.70	21:43.80	22:00.40	43:44.20
3.	4.	Chris Timms (251)	Dulwich Runners	7:16:58.30	22:15.30	23:27.00	45:42.30
4.	6.	Benoit Besnier (320)	COS Coaching	7:16:17.80	23:32.70	24:02.80	47:35.50
5.	7.	Daniel Rose (311)	Coached Fitness	7:16:59.20	22:55.10	24:41.30	47:36.40
6.	10.	Stanislav Miroshnichenko		7:18:44.00	24:22.60	24:23.00	48:45.60
7.	12.	Ian Stewart (253)	Dulwich Runners	7:17:55.80	23:27.30	25:37.30	49:04.60
8.	14.	Tycen Bundgaard (297)		7:18:43.60	25:01.40	24:47.10	49:48.50
9.	15.	Samuel Iii Belandres	Filipino Runners	7:17:55.20	23:38.40	26:18.70	49:57.10
10.	16.	Anish Jha (299)		7:17:55.70	25:28.60	25:40.90	51:09.50
11.	17.	Jonathan Eudeline (315)		7:18:44.80	25:32.60	26:09.20	51:41.80
12.	18.	Adrien Humbert (326)	Shelfordians	7:19:35.20	24:59.30	26:56.80	51:56.10
13.	19.	Ludovic Guyonvarch		7:19:35.80	25:54.60	26:59.00	52:53.60
14.	21.	Hamish Livingstone (330)		7:22:45.20	26:13.00	27:51.50	54:04.50
15.	22.	Finlay Reid (312)		7:17:55.10	25:20.00	29:18.10	54:38.10
16.	23.	Oscar Fourie (314)		8:26:19.40	26:57.20	27:42.70	54:39.90

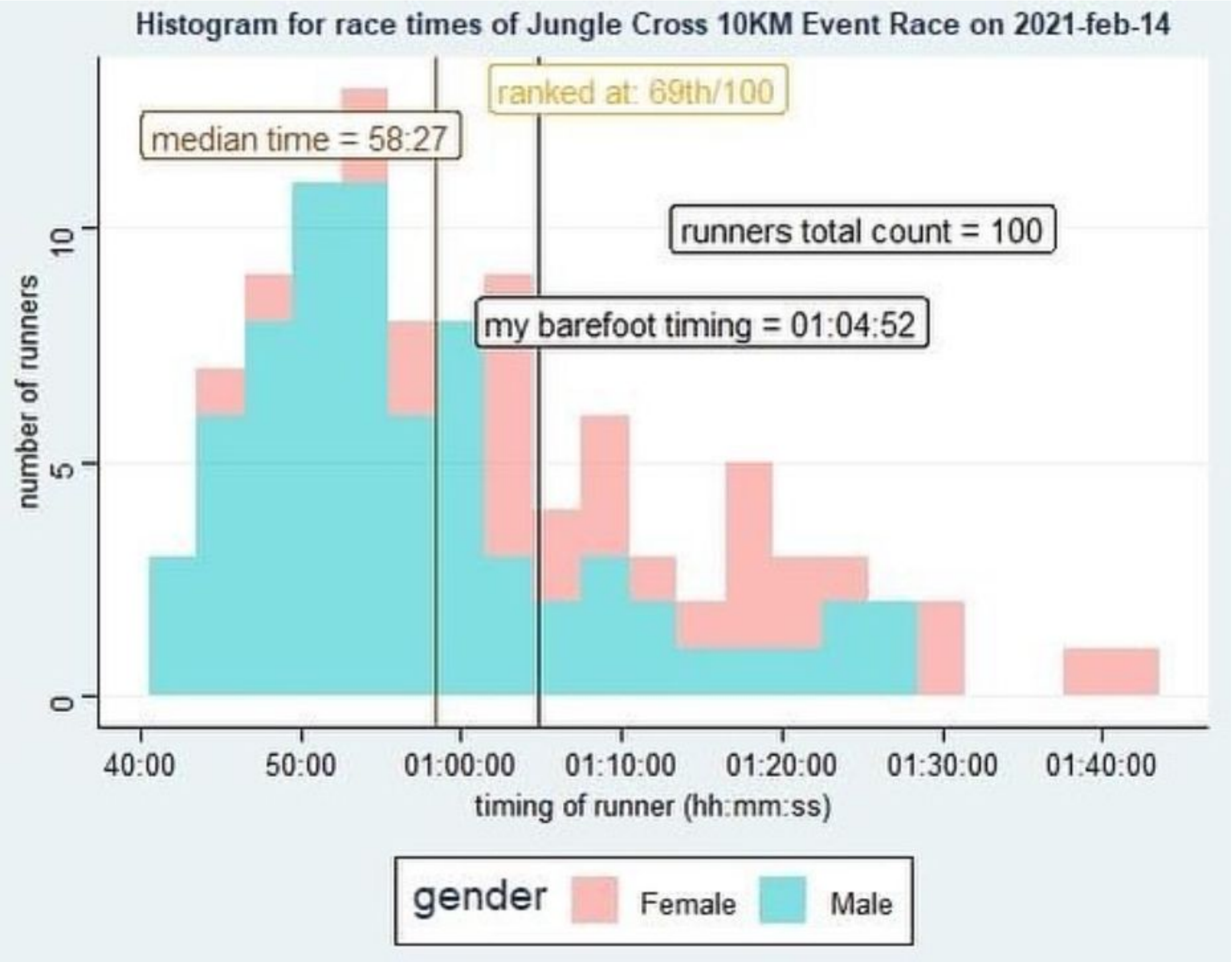
Sports analytics

How can we leverage on data to improve performance?

“Sports analytics are a collection of relevant, historical, statistics that can provide a competitive advantage to a team or individual.”

- Sports Analytics, Wikipedia

Sample Deliverable



The problem

Who was I?

An aspiring data scientist/analyst wanting to take up the challenge of Exploratory Data Analysis using Python.

What was the problem?

My friend needed to obtain useful insights about his performance in his recent trail run race.

This would be in the form of clear and effective visualisations.

Who was the audience?

My friend, a trail runner, coding primarily in R.

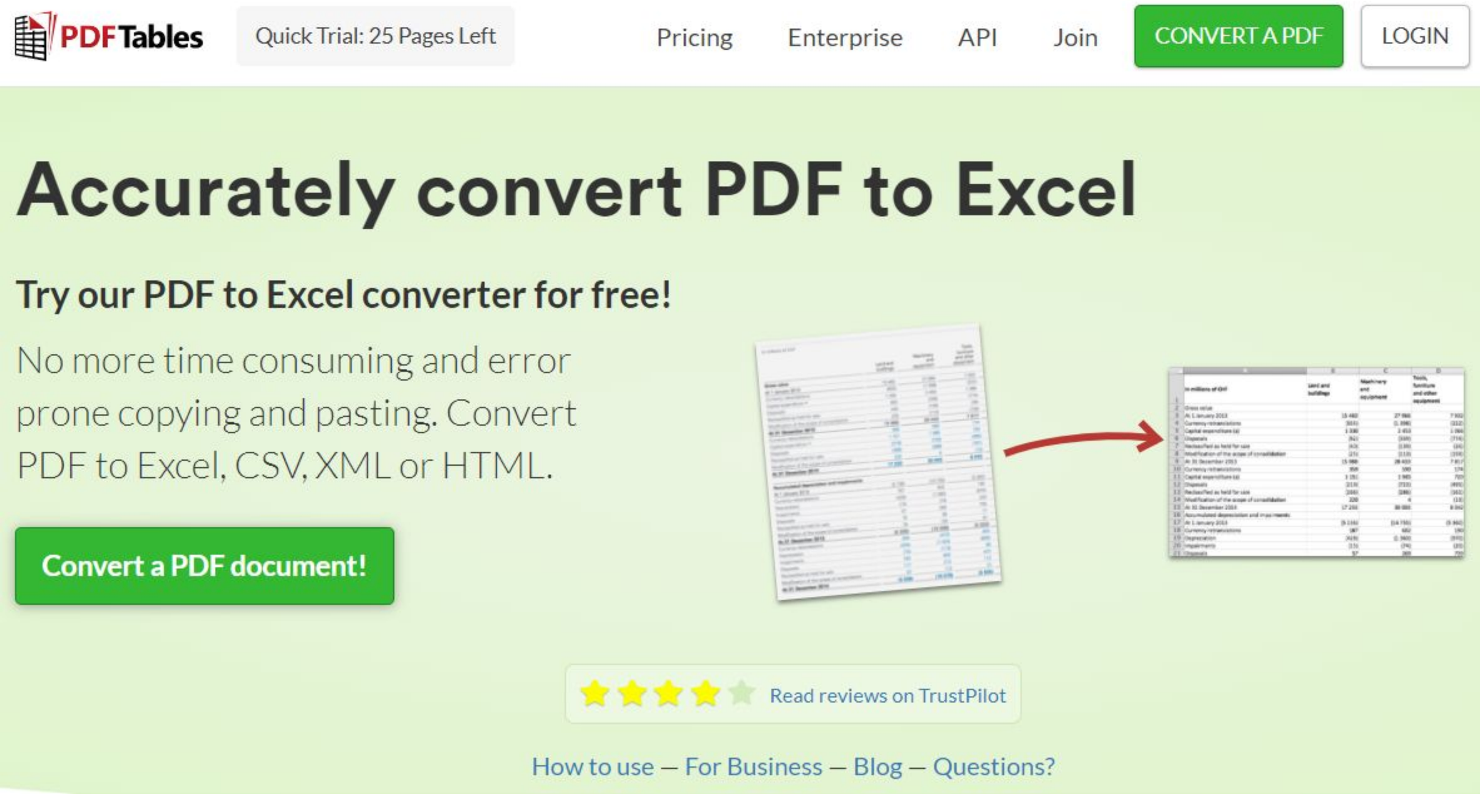
Thus, explanations should preferably be code-free.

Data Cleaning

Data Exploration

Concluding Insights

Conversion to .csv


PDFTables
Quick Trial: 25 Pages Left
Pricing
Enterprise
API
Join
CONVERT A PDF
LOGIN

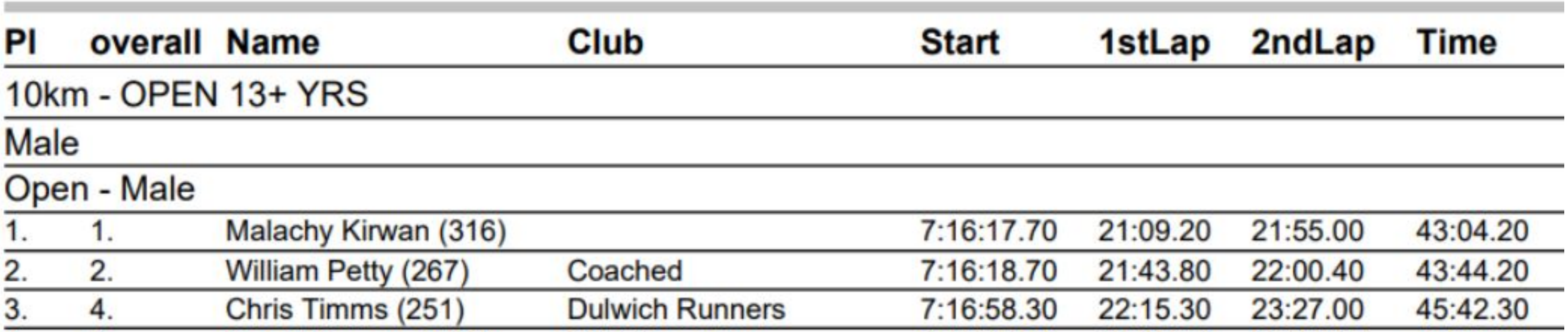
Accurately convert PDF to Excel

Try our PDF to Excel converter for free!

No more time consuming and error prone copying and pasting. Convert PDF to Excel, CSV, XML or HTML.

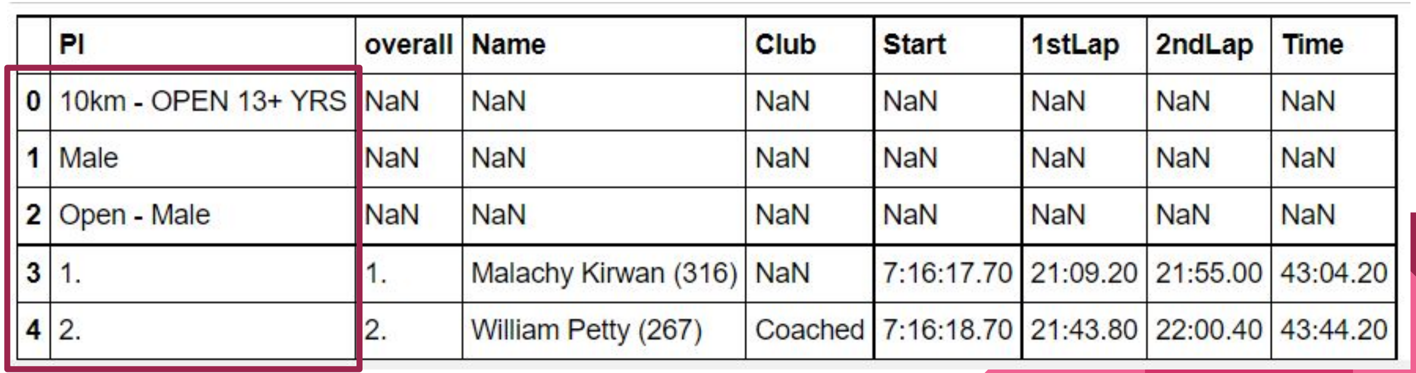
Convert a PDF document!
Read reviews on TrustPilot
How to use — For Business — Blog — Questions?

Working with the data



Pl	overall	Name	Club	Start	1stLap	2ndLap	Time
10km - OPEN 13+ YRS							
Male							
Open - Male							
1.	1.	Malachy Kirwan (316)		7:16:17.70	21:09.20	21:55.00	43:04.20
2.	2.	William Petty (267)	Coached	7:16:18.70	21:43.80	22:00.40	43:44.20
3.	4.	Chris Timms (251)	Dulwich Runners	7:16:58.30	22:15.30	23:27.00	45:42.30

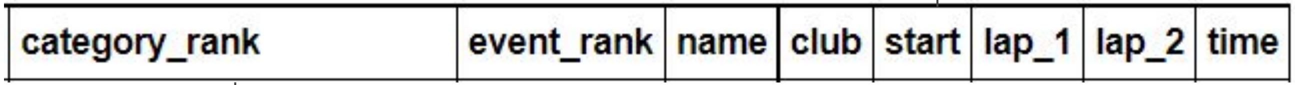
?



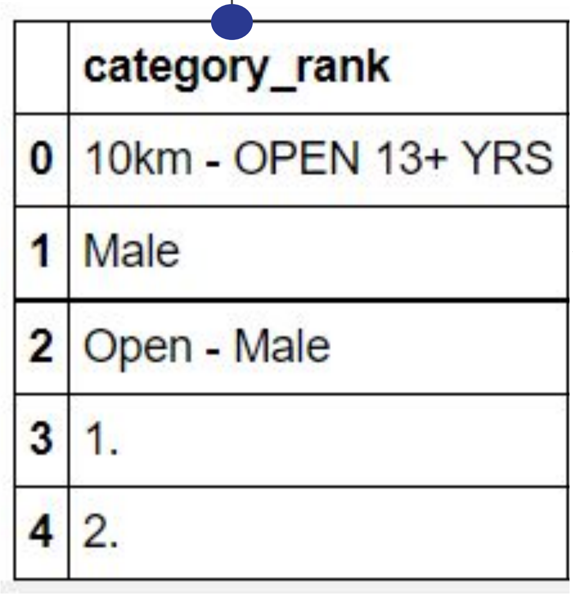
	Pl	overall	Name	Club	Start	1stLap	2ndLap	Time
0	10km - OPEN 13+ YRS	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Male	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	Open - Male	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	1.	1.	Malachy Kirwan (316)	NaN	7:16:17.70	21:09.20	21:55.00	43:04.20
4	2.	2.	William Petty (267)	Coached	7:16:18.70	21:43.80	22:00.40	43:44.20

category_rank column

Column renaming



category_rank	event_rank	name	club	start	lap_1	lap_2	time



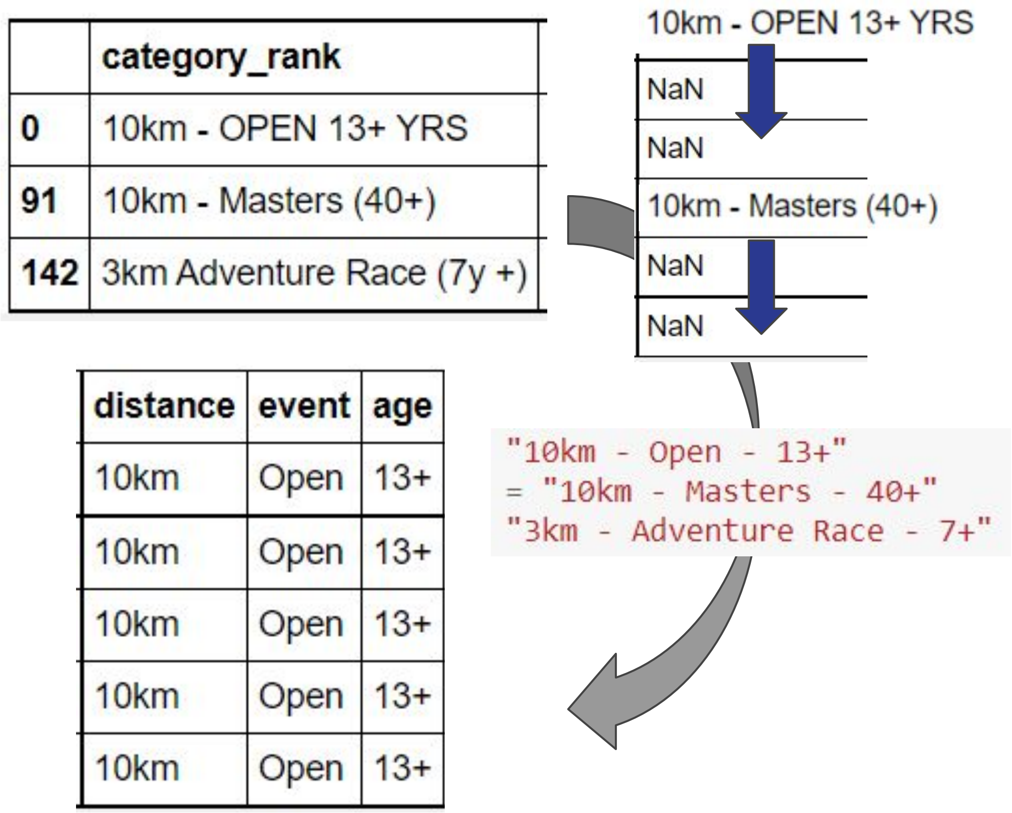
	category_rank
0	10km - OPEN 13+ YRS
1	Male
2	Open - Male
3	1.
4	2.

- Distance data (10KM)
- Age data (13+ years)
- Event data (Open)

- Gender (Male)

- Ranking data

Distance, Event and Age



	category_rank
0	10km - OPEN 13+ YRS
91	10km - Masters (40+)
142	3km Adventure Race (7y +)

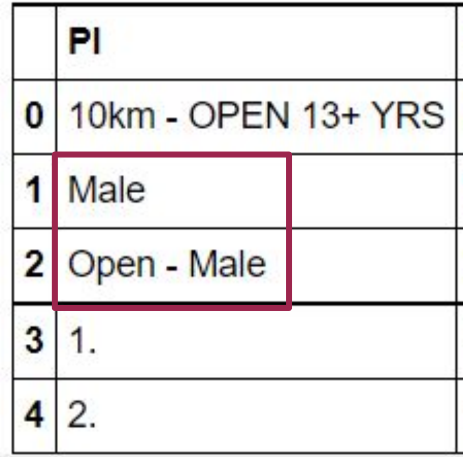
10km - OPEN 13+ YRS
NaN
NaN
10km - Masters (40+)
NaN
NaN

```
"10km - Open - 13+"
= "10km - Masters - 40+"
"3km - Adventure Race - 7+"
```

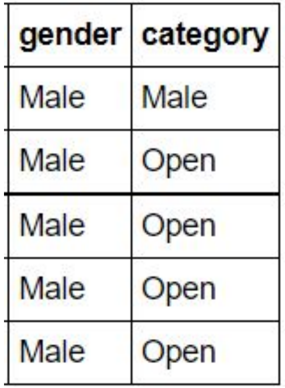
distance	event	age
10km	Open	13+
10km	Open	13+
10km	Open	13+
10km	Open	13+
10km	Open	13+

- Copied the column
- Changed all rows without “km” to a null value (NaN)
- Forward-filling: Using any valid values from above to overwrite null values going down the column
- Changed the text wording
- Split the column using dash (-) as separator
- Named new columns

Gender and Category

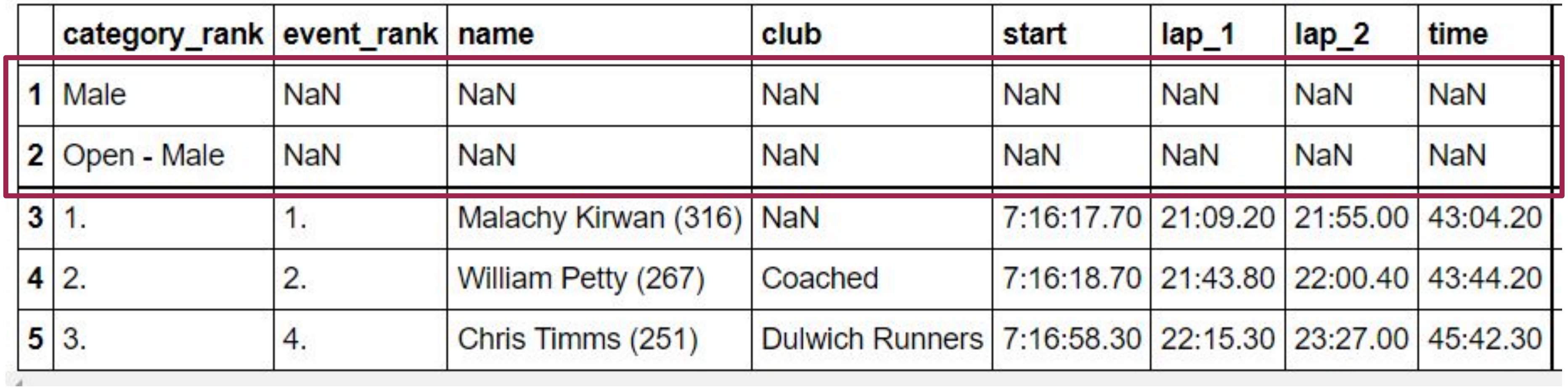


	Pl
0	10km - OPEN 13+ YRS
1	Male
2	Open - Male
3	1.
4	2.



gender	category
Male	Male
Male	Open
Male	Open
Male	Open
Male	Open

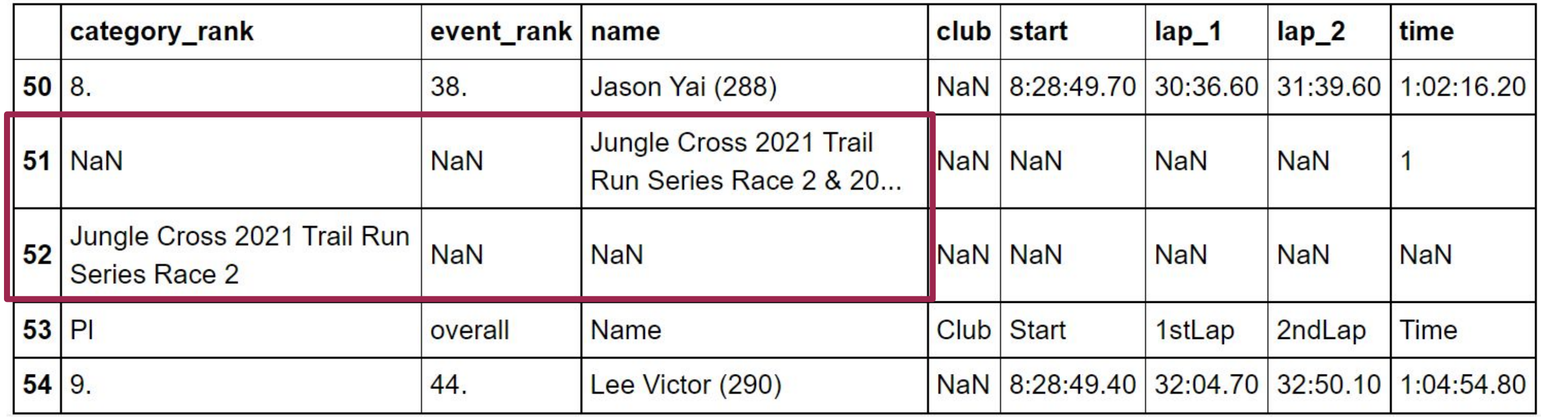
Cleaning up



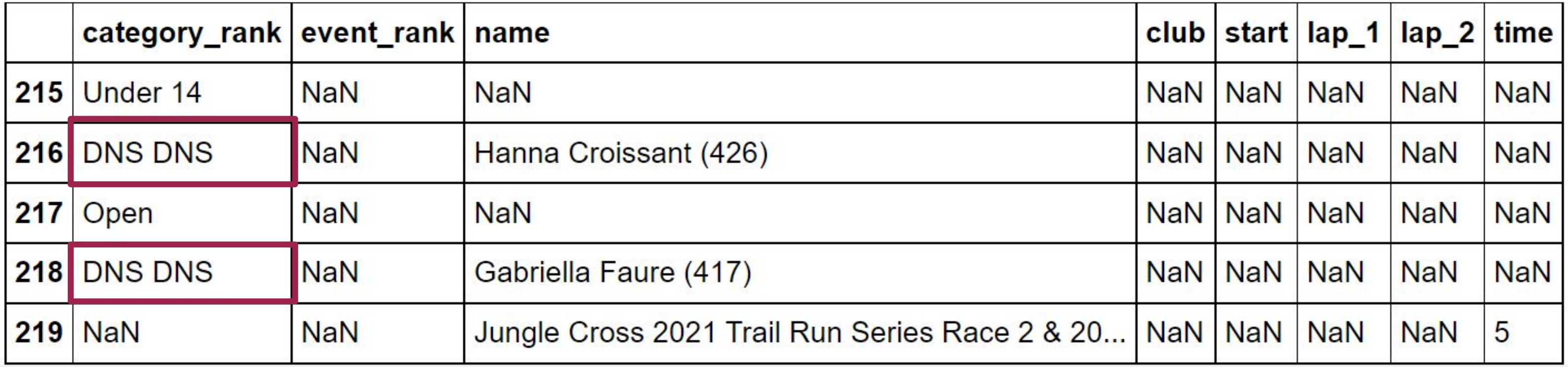
	category_rank	event_rank	name	club	start	lap_1	lap_2	time
1	Male	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	Open - Male	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	1.	1.	Malachy Kirwan (316)	NaN	7:16:17.70	21:09.20	21:55.00	43:04.20
4	2.	2.	William Petty (267)	Coached	7:16:18.70	21:43.80	22:00.40	43:44.20
5	3.	4.	Chris Timms (251)	Dulwich Runners	7:16:58.30	22:15.30	23:27.00	45:42.30

- Simply remove the rows with no time

More cleaning up

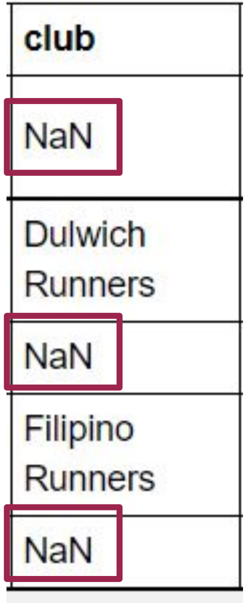


	category_rank	event_rank	name	club	start	lap_1	lap_2	time
50	8.	38.	Jason Yai (288)	NaN	8:28:49.70	30:36.60	31:39.60	1:02:16.20
51	NaN	NaN	Jungle Cross 2021 Trail Run Series Race 2 & 20...	NaN	NaN	NaN	NaN	1
52	Jungle Cross 2021 Trail Run Series Race 2	NaN	NaN	NaN	NaN	NaN	NaN	NaN
53	Pl	overall	Name	Club	Start	1stLap	2ndLap	Time
54	9.	44.	Lee Victor (290)	NaN	8:28:49.40	32:04.70	32:50.10	1:04:54.80

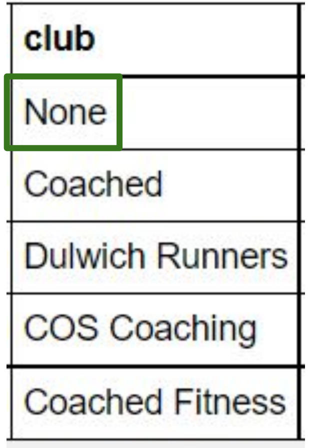


	category_rank	event_rank	name	club	start	lap_1	lap_2	time
215	Under 14	NaN	NaN	NaN	NaN	NaN	NaN	NaN
216	DNS DNS	NaN	Hanna Croissant (426)	NaN	NaN	NaN	NaN	NaN
217	Open	NaN	NaN	NaN	NaN	NaN	NaN	NaN
218	DNS DNS	NaN	Gabriella Faure (417)	NaN	NaN	NaN	NaN	NaN
219	NaN	NaN	Jungle Cross 2021 Trail Run Series Race 2 & 20...	NaN	NaN	NaN	NaN	5

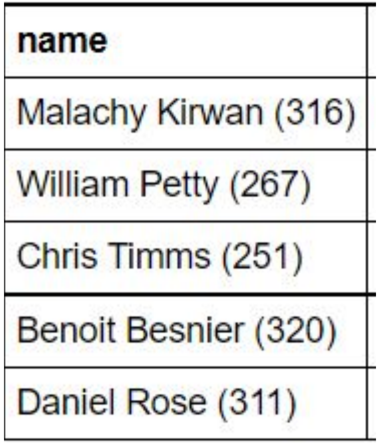
Club column and Bib Number column



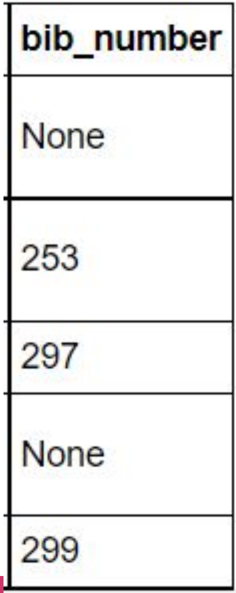
club
NaN
Dulwich Runners
NaN
Filipino Runners
NaN



club
None
Coached
Dulwich Runners
COS Coaching
Coached Fitness

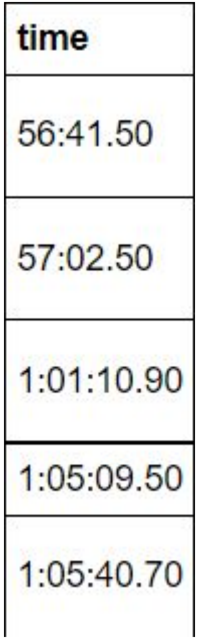


name
Malachy Kirwan (316)
William Petty (267)
Chris Timms (251)
Benoit Besnier (320)
Daniel Rose (311)



bib_number
None
253
297
None
299

Timedelta object



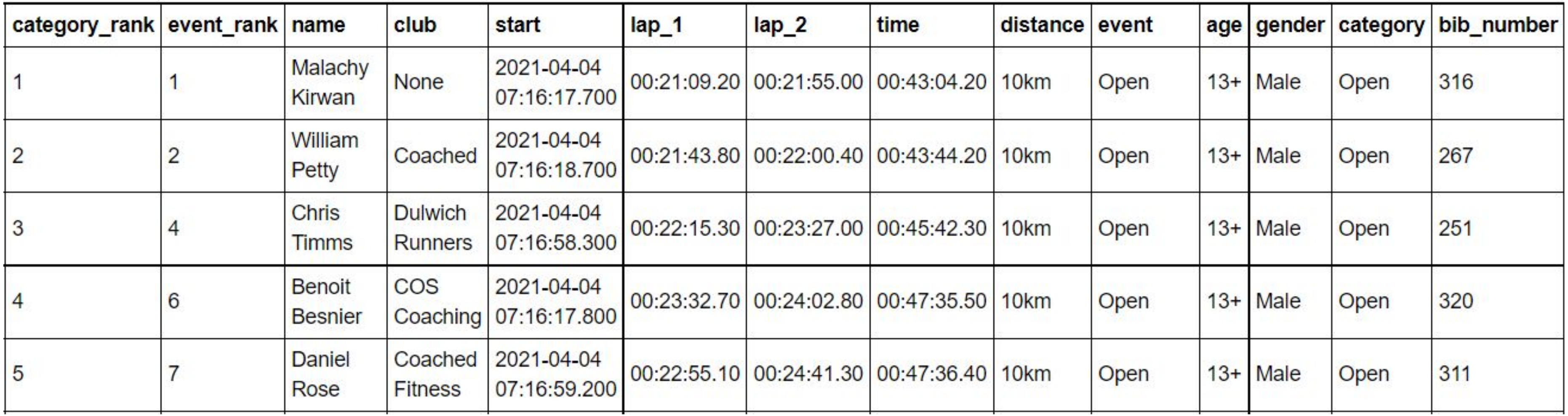
time
56:41.50
57:02.50
1:01:10.90
1:05:09.50
1:05:40.70

- Some were in MM:SS and some were in H:MM:SS
- Could not convert to timedelta object (for analysis)
- Want everything in HH:MM:SS
- We need to add extra ‘**00:**’s to those below an hour
- We need to add extra ‘**0:**’s to those one hour and above

Solutions (to differentiate them):

- Use number of characters
- Use number of semicolons
- Use Regex

Cleaned data



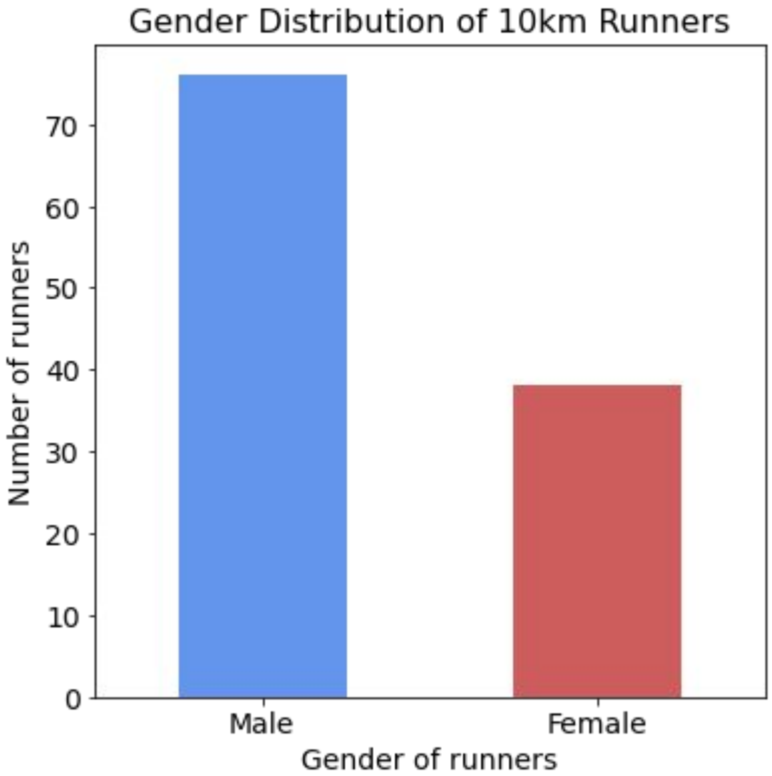
category_rank	event_rank	name	club	start	lap_1	lap_2	time	distance	event	age	gender	category	bib_number
1	1	Malachy Kirwan	None	2021-04-04 07:16:17.700	00:21:09.20	00:21:55.00	00:43:04.20	10km	Open	13+	Male	Open	316
2	2	William Petty	Coached	2021-04-04 07:16:18.700	00:21:43.80	00:22:00.40	00:43:44.20	10km	Open	13+	Male	Open	267
3	4	Chris Timms	Dulwich Runners	2021-04-04 07:16:58.300	00:22:15.30	00:23:27.00	00:45:42.30	10km	Open	13+	Male	Open	251
4	6	Benoit Besnier	COS Coaching	2021-04-04 07:16:17.800	00:23:32.70	00:24:02.80	00:47:35.50	10km	Open	13+	Male	Open	320
5	7	Daniel Rose	Coached Fitness	2021-04-04 07:16:59.200	00:22:55.10	00:24:41.30	00:47:36.40	10km	Open	13+	Male	Open	311

Data Cleaning

Data Exploration

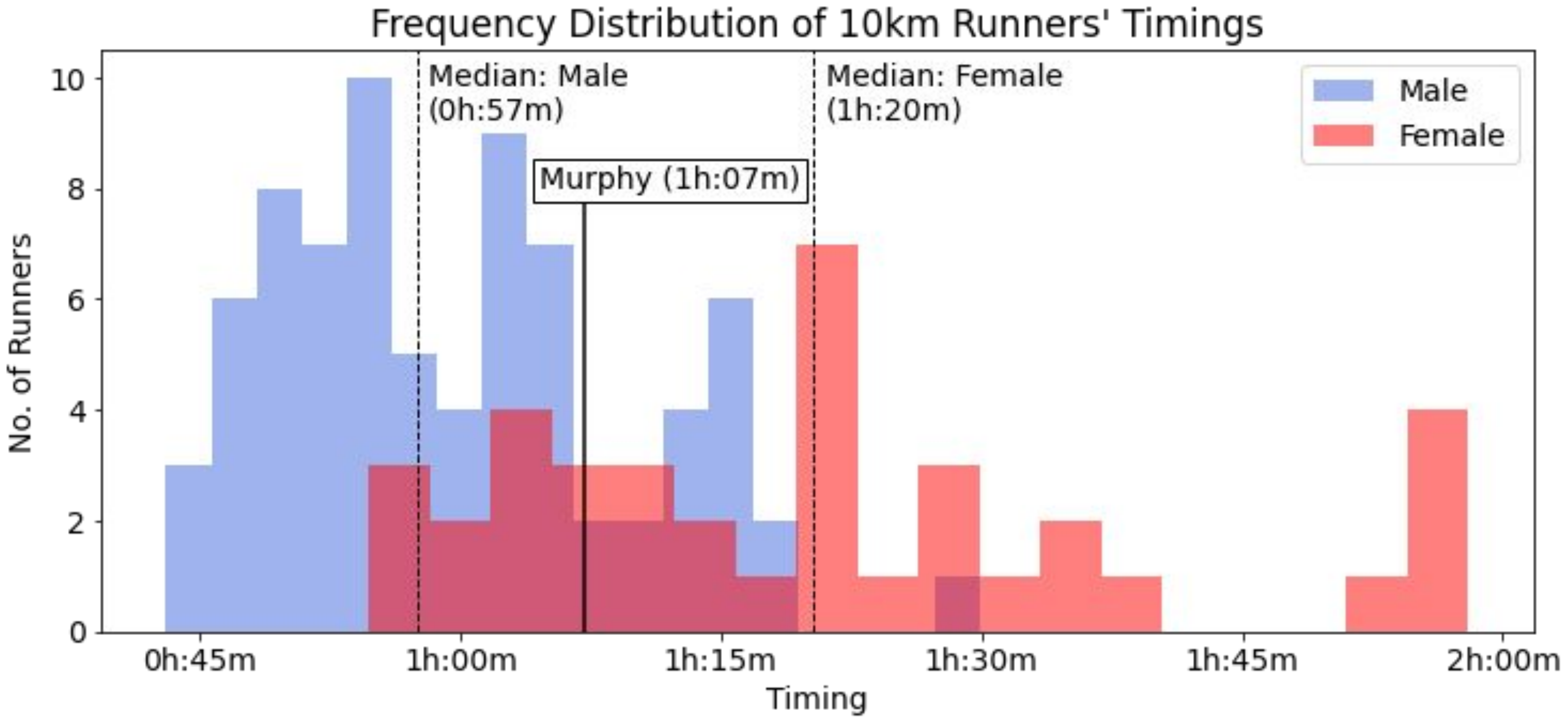
Concluding Insights

Gender

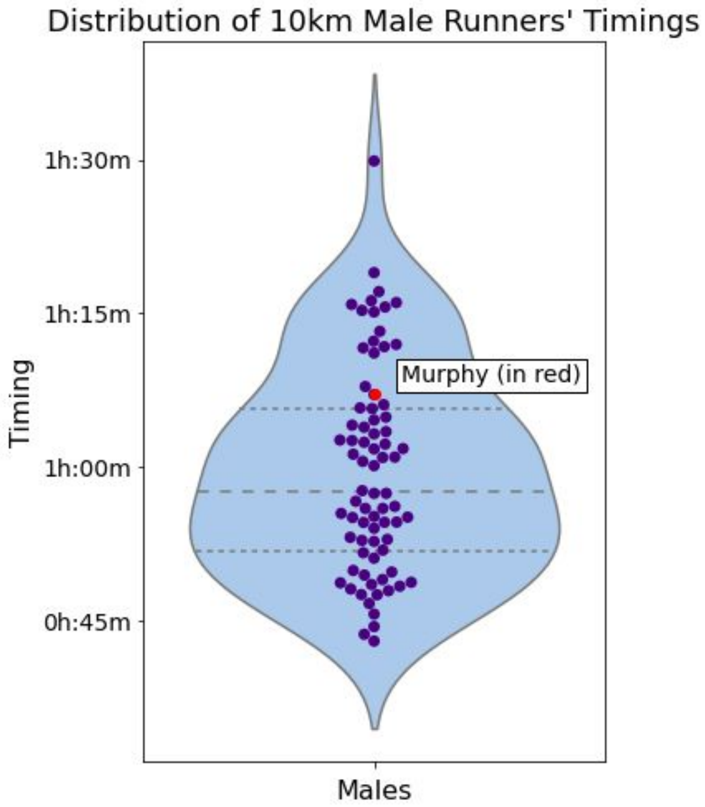


- Only 10KM runners were analysed
- 114 runners in data

Visualising the distribution

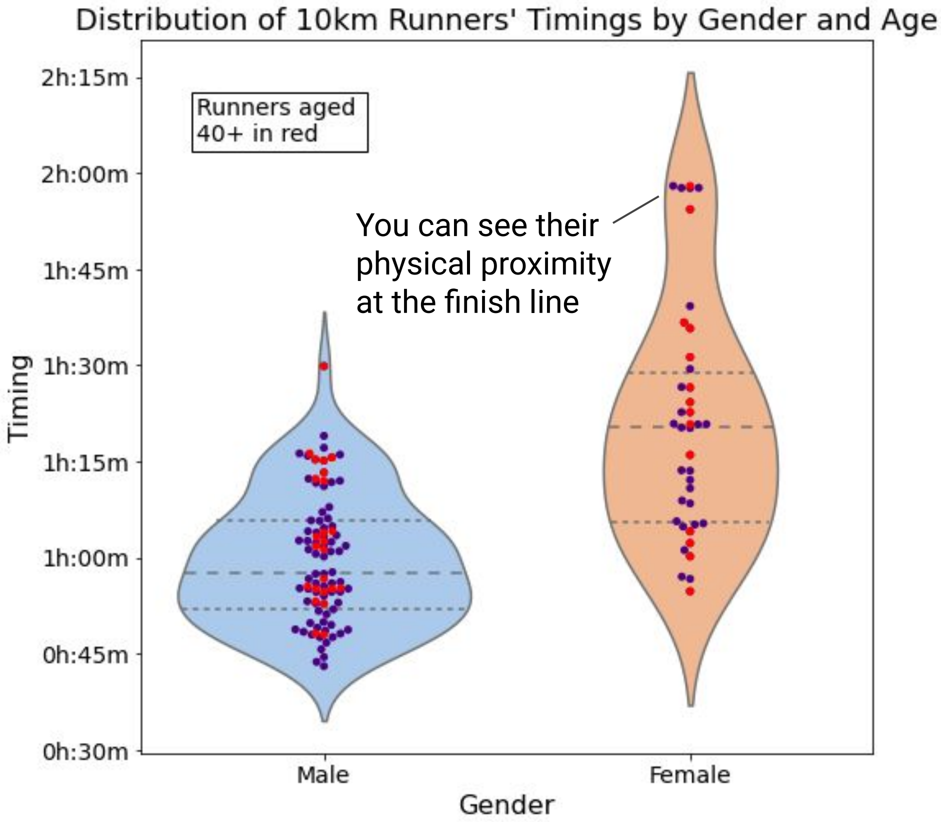


Highlighting an individual



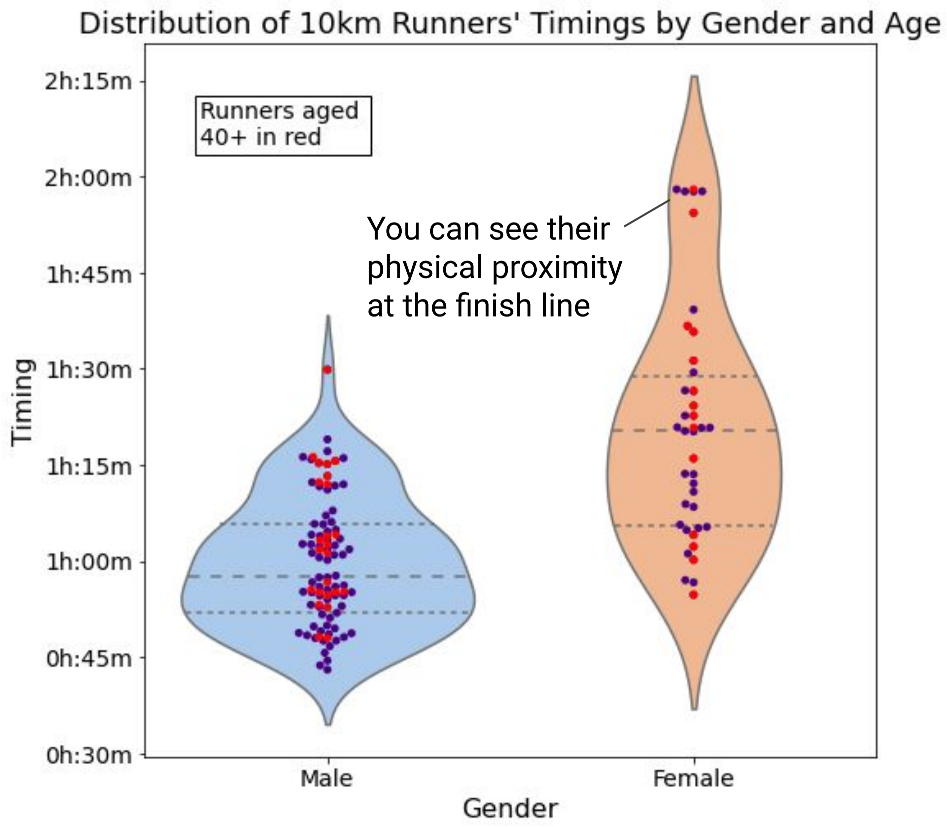
- Swarm plot overlaid on a violin plot
- Dotted lines are quartiles
- Now we can see that the timing is behind the 75th percentile

Gender and Age



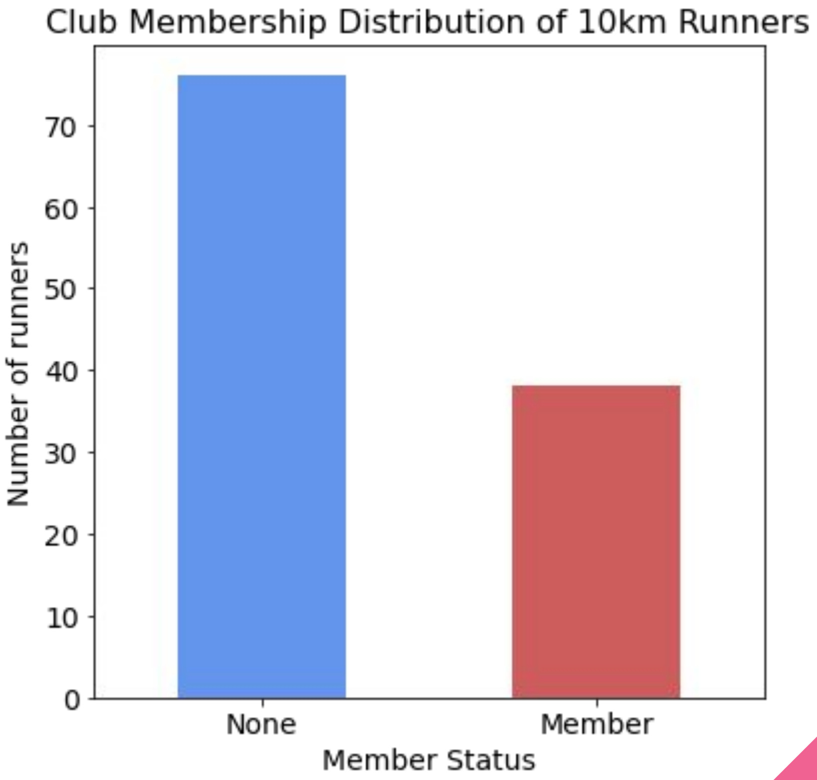
You can see their physical proximity at the finish line

Gender and Age

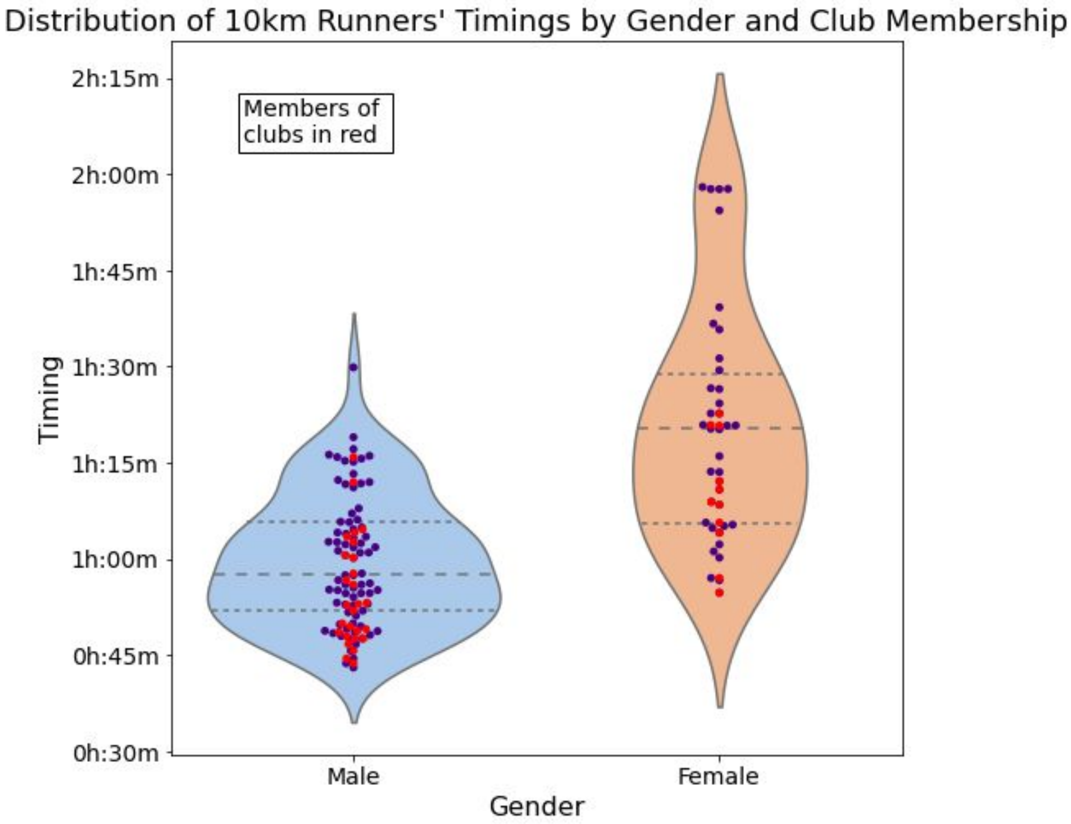


- Median for males 40+: 1h:02m
- Median for males 13+: **0h:56m**
- Median for females 40+: 1h:23m
- Median for females 13+: **1h:13m**

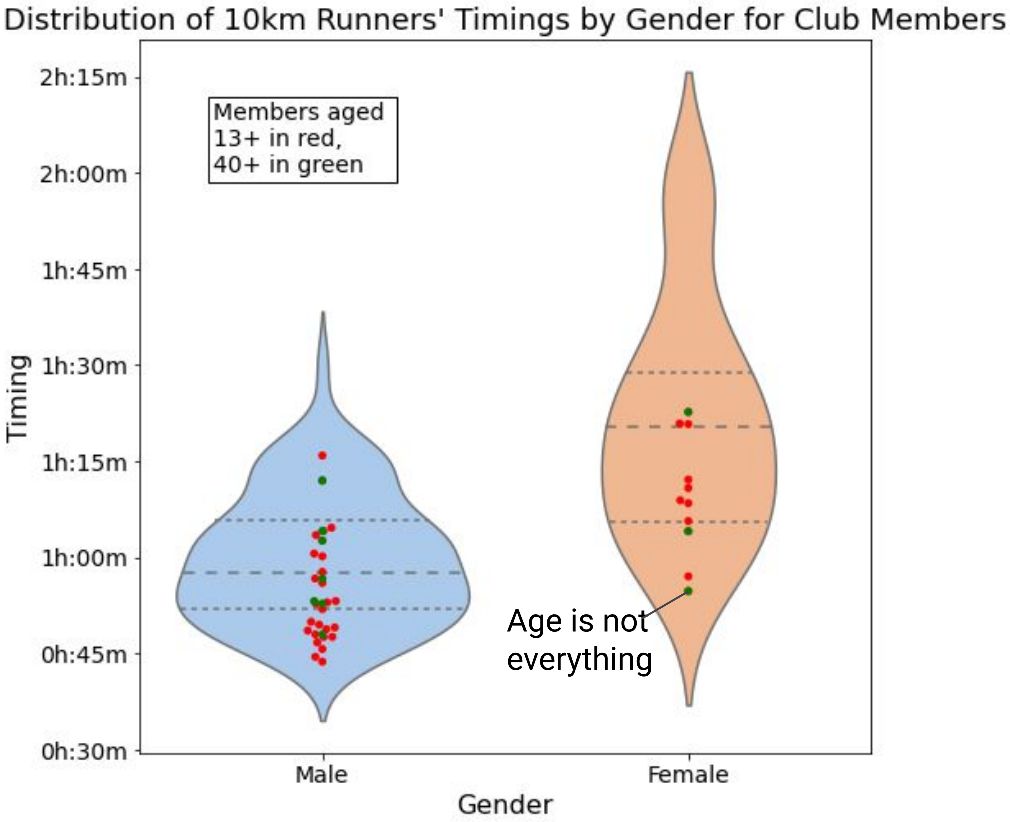
Club Membership



Gender and Club Membership

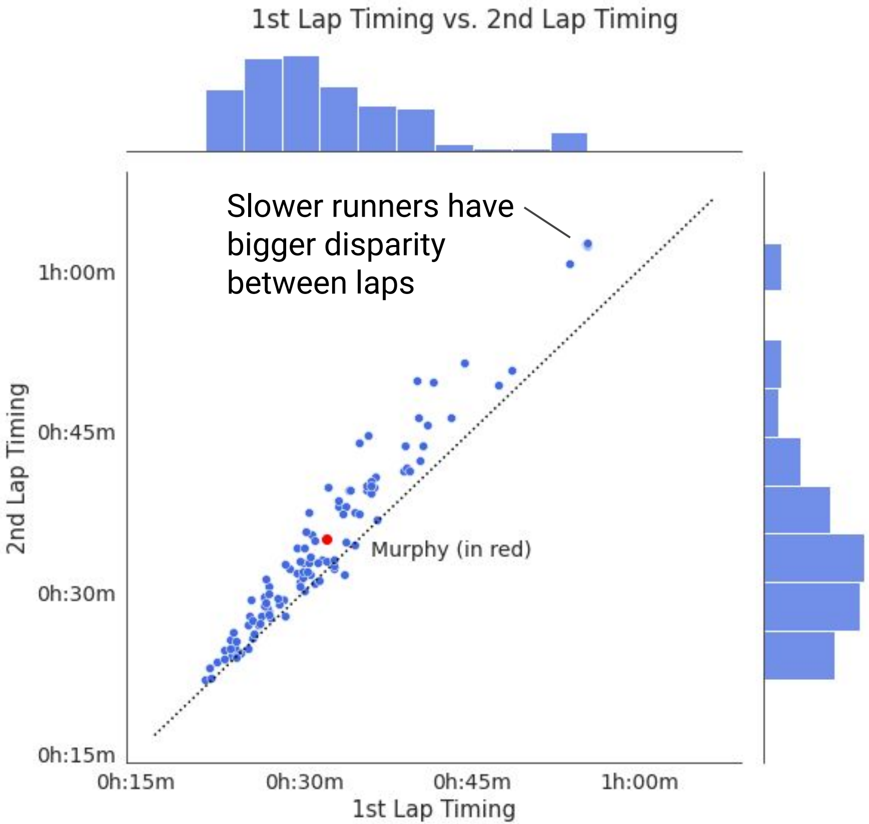


Gender and Age of Club Members only



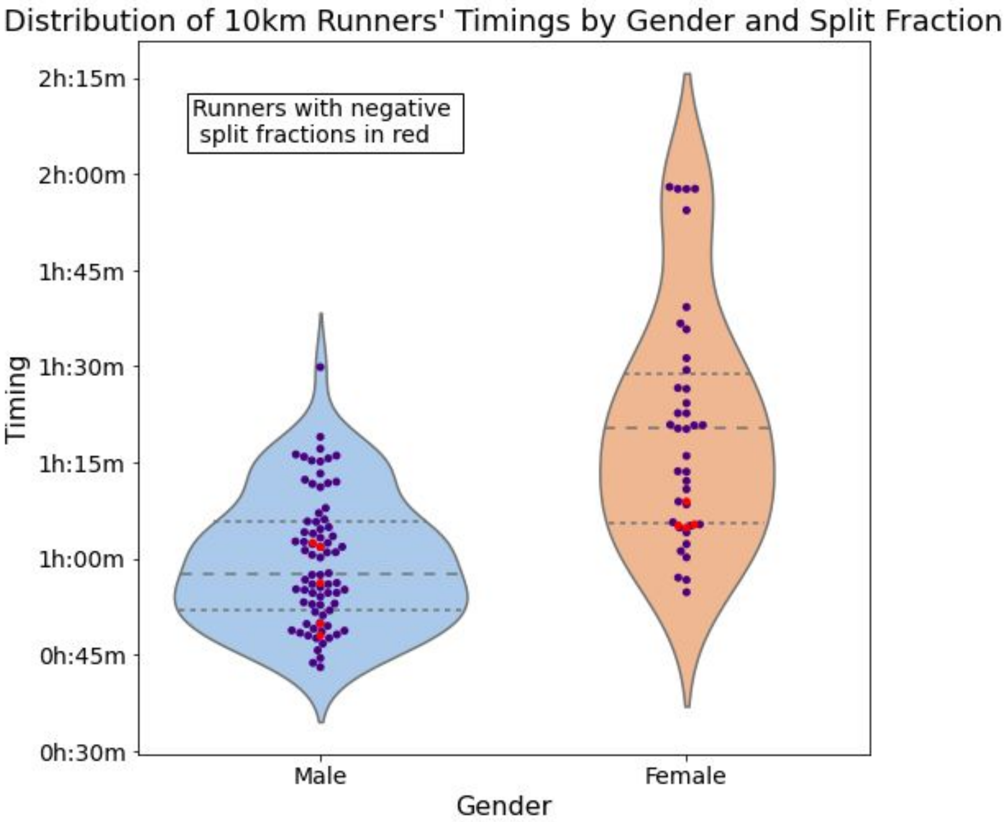
Age is not everything

Race Split (1st Lap vs. 2nd Lap)



- Normal to take more time for 2nd lap
- 9 out of 114 runners ran 2nd lap faster
- This is 7.9% of runners

Gender and Race Split



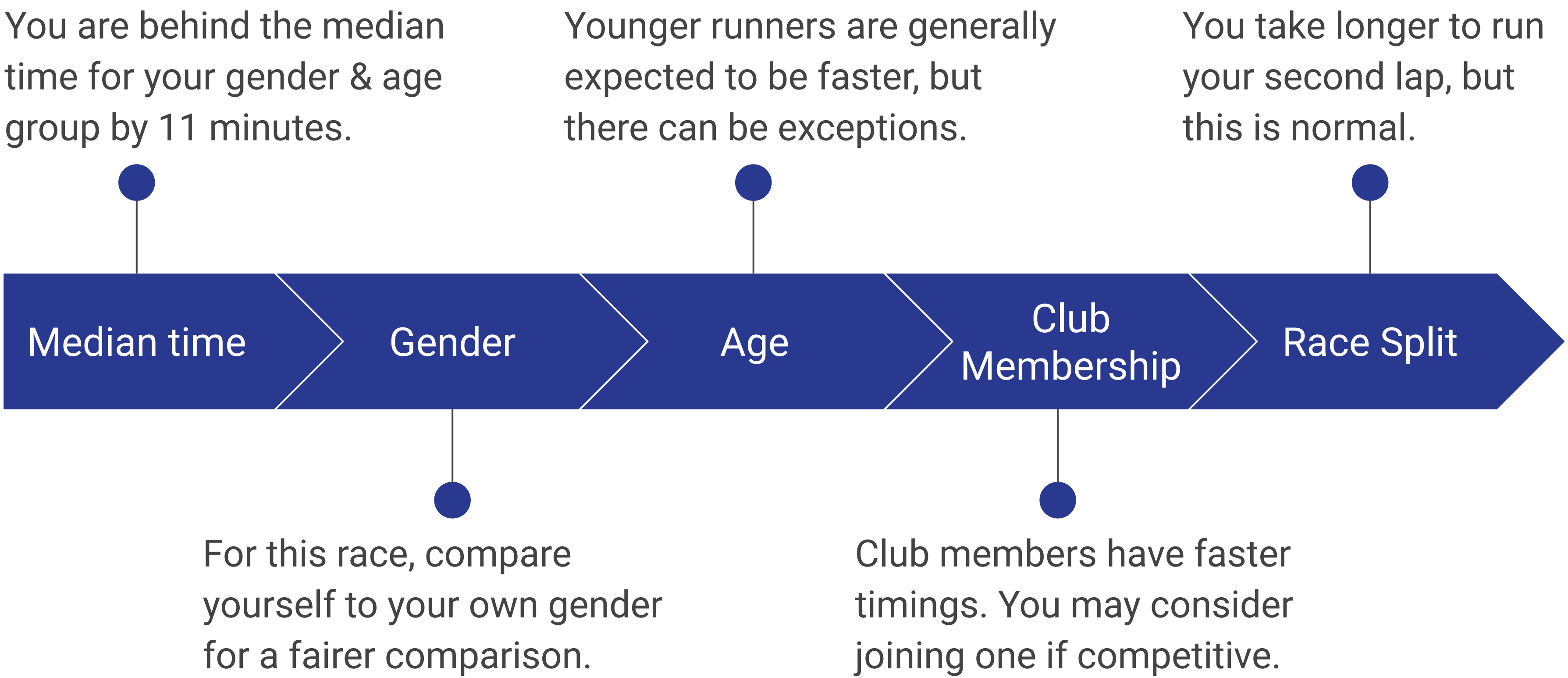
Split Fraction =
1 - (1st Lap)/(2nd Lap)

Data Cleaning

Data Exploration

Concluding Insights

Personalised Analytics Example



Median time

You are behind the median time for your gender & age group by 11 minutes.

Gender

For this race, compare yourself to your own gender for a fairer comparison.

Age

Younger runners are generally expected to be faster, but there can be exceptions.

Club Membership

Club members have faster timings. You may consider joining one if competitive.

Race Split

You take longer to run your second lap, but this is normal.

Key Takeaways from Project

Value of EDA

Exploratory Data Analysis already reveals key insights

Scouted online for ideas on race analysis with Python

Swarm plots

Data Science Problems

They are all around you

Talk to the people around you

Different people have different receptivity/ resistance

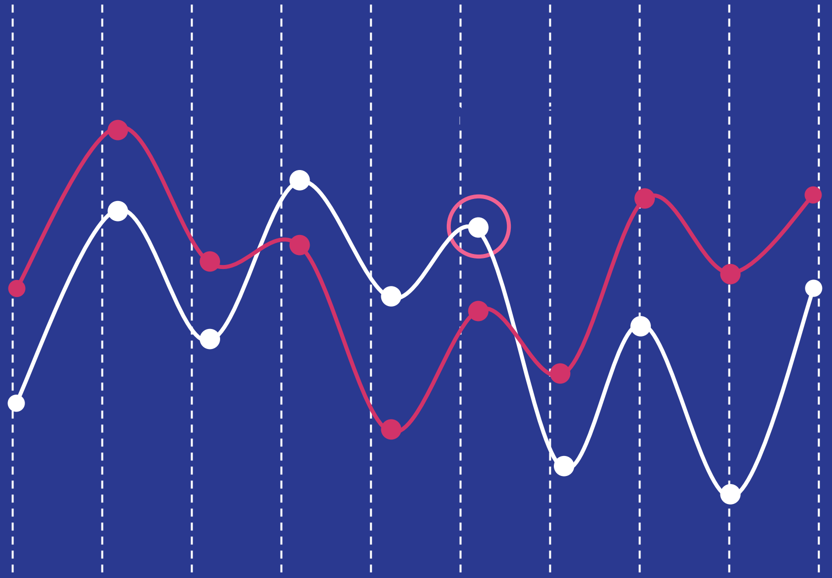
Project-based learning

Just need to go through the fire initially

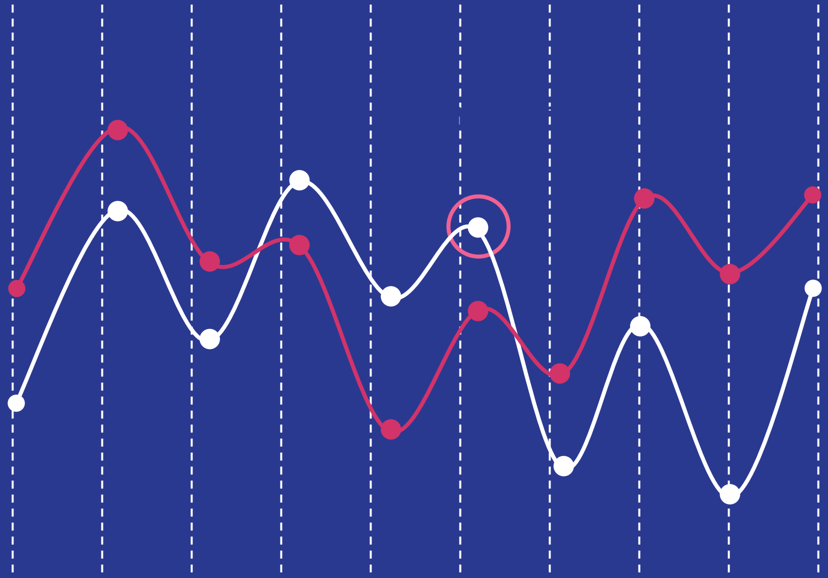
How many times do you have to relearn Pandas (for e.g.) in your life?

It will become easier

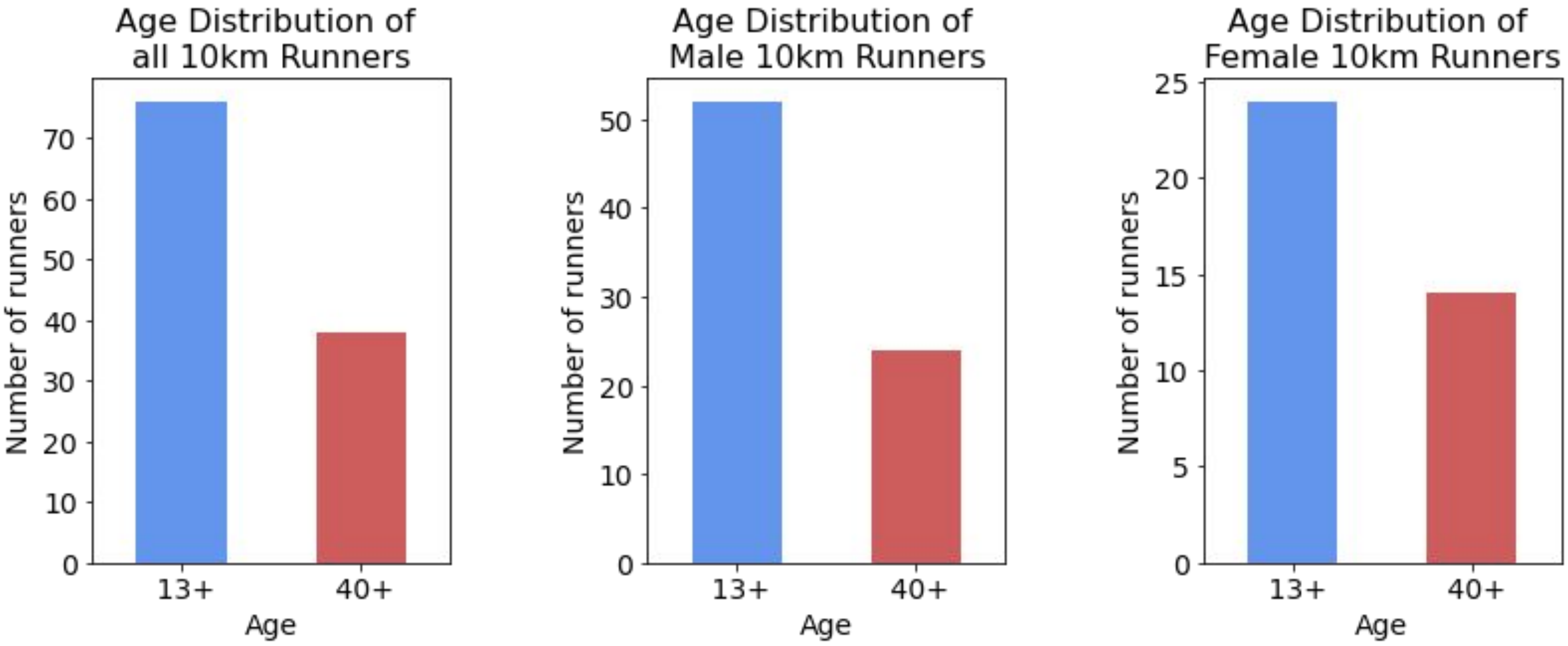
Thank you!



Q&A



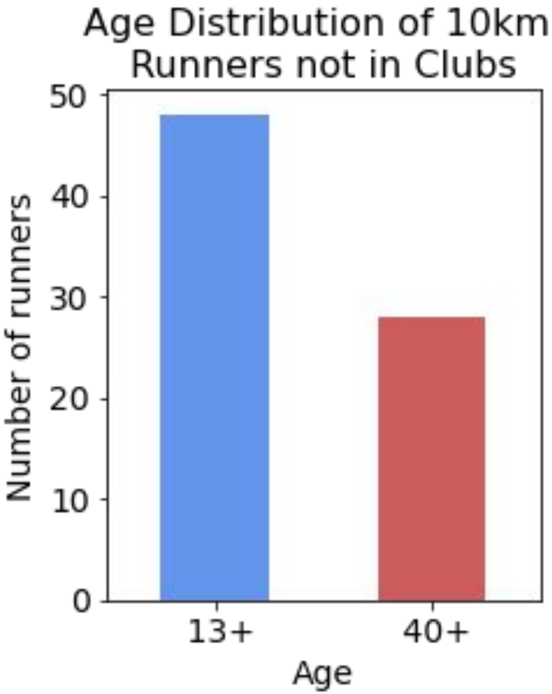
Age distributions


Age Distribution of all 10km Runners

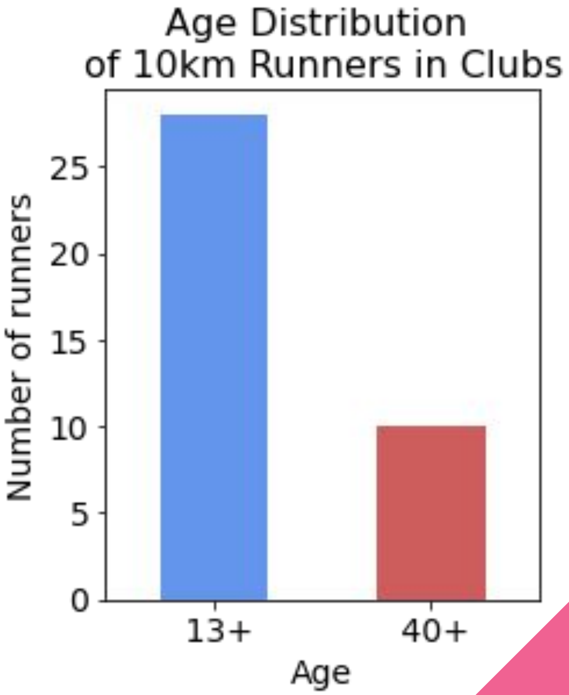
Age Distribution of Male 10km Runners

Age Distribution of Female 10km Runners

Age and Club Membership

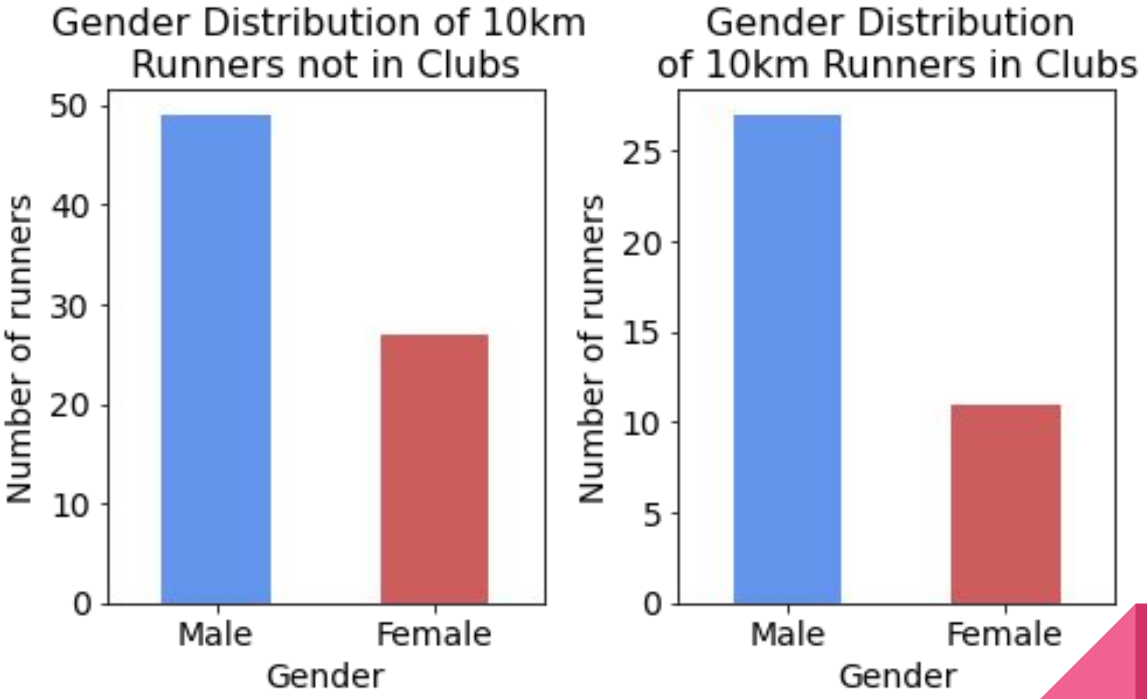


Age Distribution of 10km Runners not in Clubs



Age Distribution of 10km Runners in Clubs

Gender and Club Membership



Gender, Age, Club Membership

- Median for males 13+ **NOT in clubs**: 1h:01m
- Median for males 13+ in clubs: **0h:49m**
- Median for males 40+ **NOT in clubs**: 1h:03m
- Median for males 40+ in clubs: **0h:56m**
- Median for females 13+ **NOT in clubs**: 1h:20m
- Median for females 13+ in clubs: **1h:09m**
- Median for females 40+ **NOT in clubs**: 1h:26m
- Median for females 40+ in clubs: **1h:04m**