Dashathon

Marathon Training Tool for Runners

Annual Marathons: a brief history

The **Abbott World Marathon Majors** is a championship-style competition for marathon runners that started in 2006. A points based competition founded on six major city marathon races, the series currently comprises annual races for the cities of Tokyo, Boston, London, Berlin, Chicago and New York City.

- **Boston:** World's oldest marathon, it began in 1897. It ranks as one of the world's best-known road racing events. There were 30,088 participants in 2018.
- **Chicago**: Started in 1977, the Chicago Marathon is the fourth-largest race by number of finishers worldwide. 44,571 finishers in 2018.
- **Berlin**: Initiated in 1974, 61,390 runners participated in 2018**.** Berlin is one of four world-wide marathons with more than 40,000 finishers.
- **London**: With the first run in 1971, it had 41,003 runners in 2019.
- **New York City**: It is the largest marathon in the world. In 2018, more than 100,000 runners applied, and the final field had 50,000 runners.

Getting the data: Scraping

Web scraping is an online data acquisition method where selected info is programmatically downloaded from a website.

Most marathon websites publically post details about their participants online, including split times. We decided to use web scraping for the Chicago, London, and Berlin marathons in order to obtain more split time data and, ultimately, to produce a more credible dashboard.

Implementation:

- *Main Python packages*: mechanize for completing web forms and bs4 for parsing html
- *Constraints*: only request information at most once every second
- *Resolved technical issues*:
    - Caching retrieved data to start/stop scraping as needed
    - Connection errors
    - Managing corner cases (missing data, inconsistent formatting, ascii vs unicode)

Data Description

We have taken the finishers data* from all the mentioned Marathons over years **2013-2017**. It contains the name, age, gender, country, city, times at 9 different stages of the race, expected time, finish time and pace, overall place, gender place and division place.

- Bib number for the year is the unique identifier of the runner
- The splits that we have are: 5k, 10k, 15k, 20k, half-way, 25k, 30k, 35k, and 40k
- Age on race day and the gender, along with the age bucket of the given marathon
- Runner's overall pace
- Runner's official finishing time
- Runner's overall ranking in a given year
- Runner's ranking in their gender
- Runner's ranking in their age division

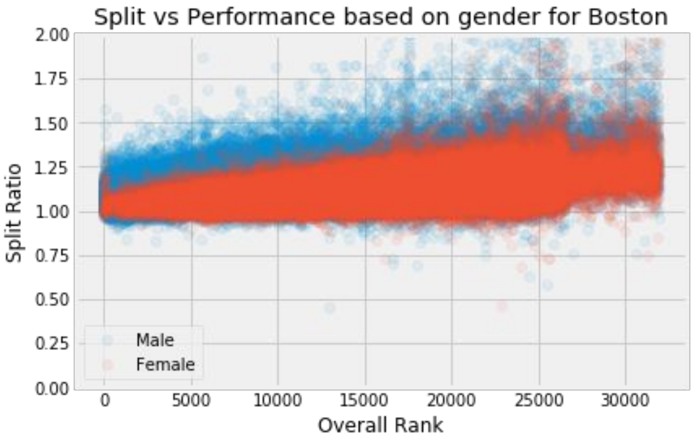
*The data in different datasets is in different formats. This is an overall structure of what we have

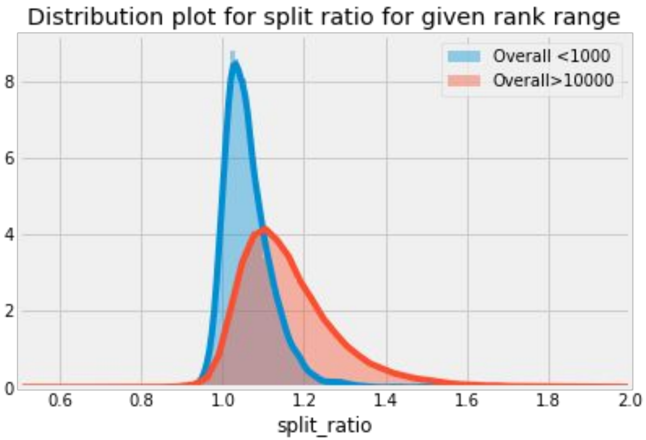
What do these features mean ?

- The splits represent the time taken (in mins) to cover the given distance in meters
- The pace is the minute per mile of the runner
- Gender Rank is the Rank in your gender and Age Rank is rank in age buckets set by us (for uniformity across datasets)
- Wall Split is the split where "The Wall" occurs - a condition of sudden fatigue which typically hits the marathon runner after about 30Ks (though of course varies among different individuals)
- Split Ratio is the ratio of time taken to run second half over first half in order to understand the negative split strategy, in which the runner runs the second part faster than the first part

  Split ratio of less than 1 represents a negative split

EDA

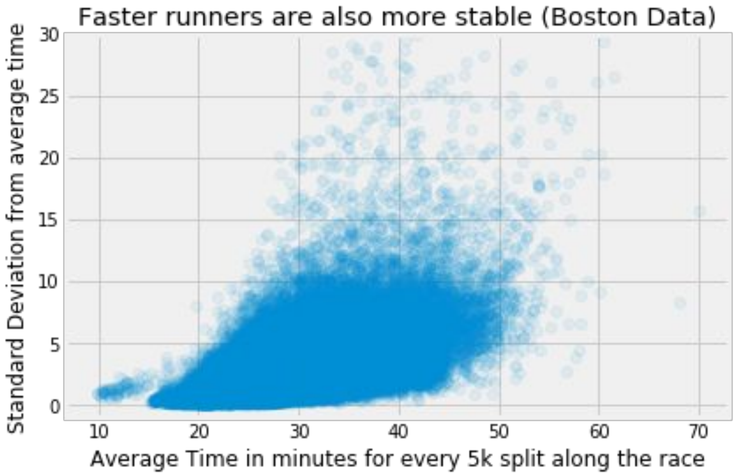


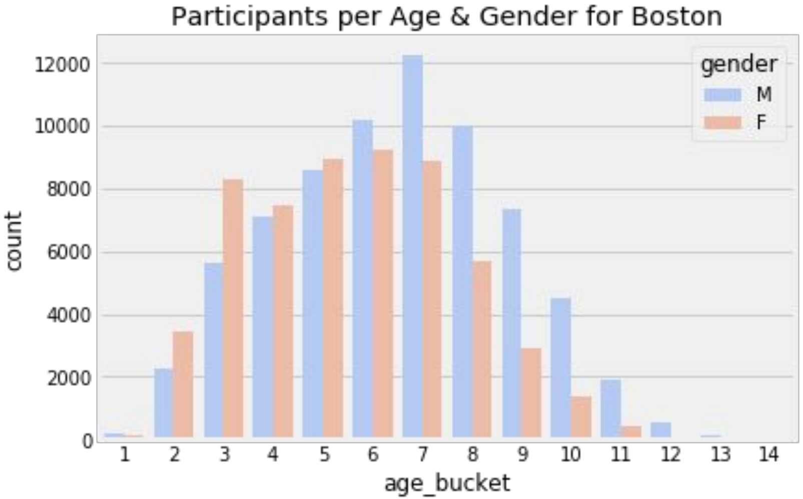


- A lot more women tend to go for negative split over men
- Increasing trend of ranks with increase in split ratio
- Variance in splits is higher as overall rank progresses- are constant pace runners doing better?

- The better runners have a less positive split than worse runners
- Runners with greater ranks tend to run a very positive split

EDA





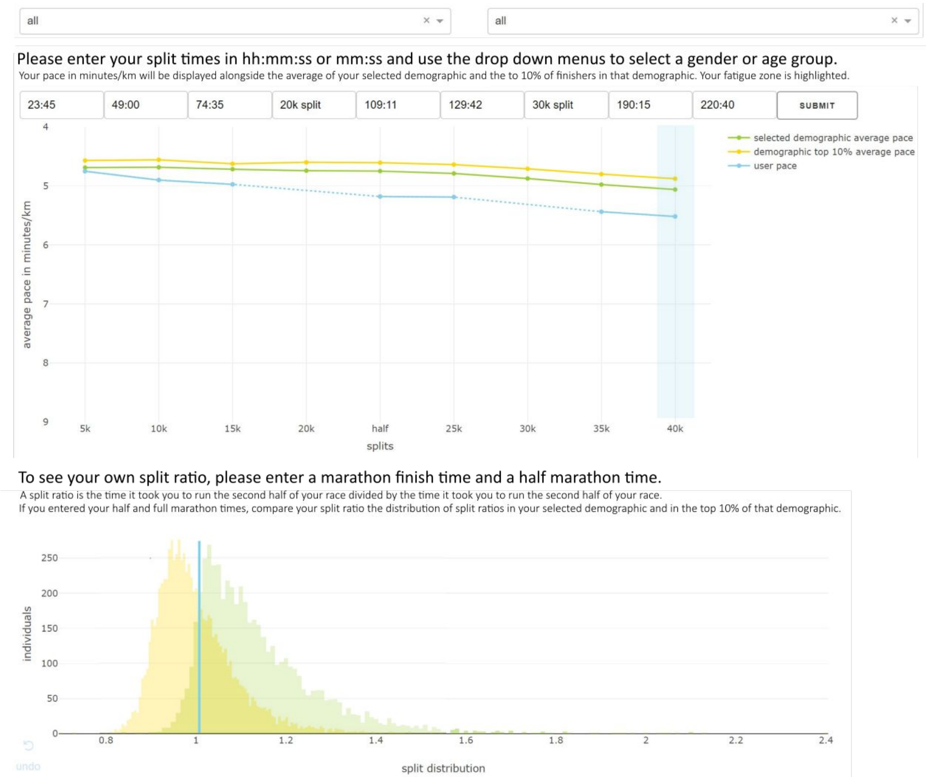
- The faster the runner (low per split time), the lower their deviation from constant pace
- Fast runners tend to not deviate too much from a constant value throughout the race

- Female finishers tend to be younger than men
- Age bucket 7 has highest male participation and highest overall participation

Merging Data

- Our current data comes from two distinct sources: Boston Marathon finishers 2013-2014 and 2015-2017.
- Basic cleaning was performed on all years and runners with unintelligible or missing finish times were dropped,
- All large race datasets do contain missing splits due to the imperfect nature of timing mats and chips. Runners with individual missing splits mid-race were not dropped.
- Mobility impaired and visually impaired runners were excluded.
- All times were converted to a base unit of seconds.
- Runners were binned based on age, using the parameters by which the Boston Athletic Association issues age group awards (these are common race age groups).
- All years were merged based on key data: gender, age, age group, split times, finish time, rank (placing), race year, etc.

Targets for the tool- ideally what we want



all

all

Please enter your split times in hh:mm:ss or mm:ss and use the drop down menus to select a gender or age group.

Your pace in minutes/km will be displayed alongside the average of your selected demographic and the to 10% of finishers in that demographic. Your fatigue zone is highlighted.

23:45 | 49:00 | 74:35 | 20k split | 109:11 | 129:42 | 30k split | 190:15 | 220:40 | SUBMIT

4
5
6
7
8
9
average pace in minutes/km
5k
10k
15k
20k
half
25k
30k
35k
40k
splits
selected demographic average pace
demographic top 10% average pace
user pace

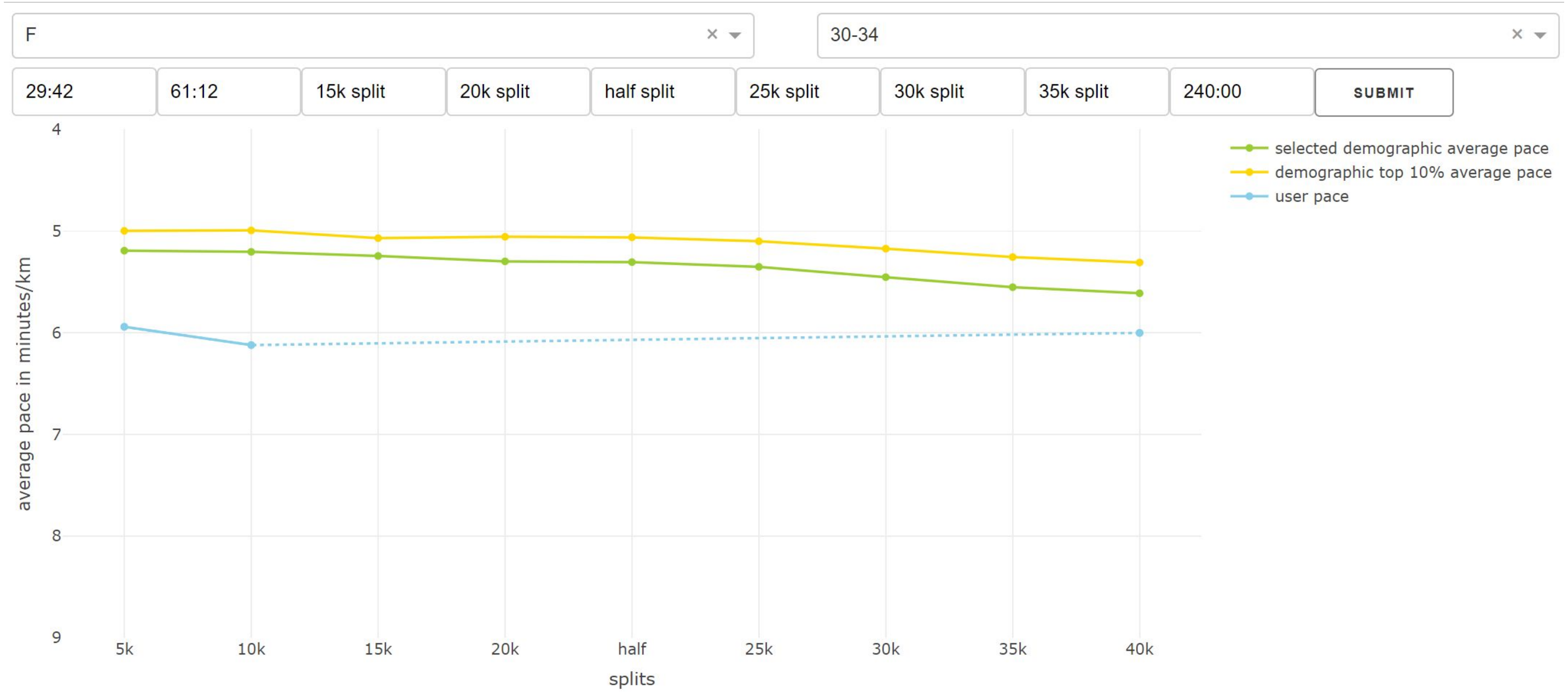
To see your own split ratio, please enter a marathon finish time and a half marathon time.

A split ratio is the time it took you to run the second half of your race divided by the time it took you to run the second half of your race.

If you entered your half and full marathon times, compare your split ratio the distribution of split ratios in your selected demographic and in the top 10% of that demographic.

- Multiple analytics
- Clear explanatory text
- Consistent design elements
- Allows user to explore data for previous finishers with or without entering their own split times, gender, or approximate age
- Offers projected finish time and/or projected rank based on user data
- Highlight the user's fatigue zone (slowest split)

Where are we right now- Current Output



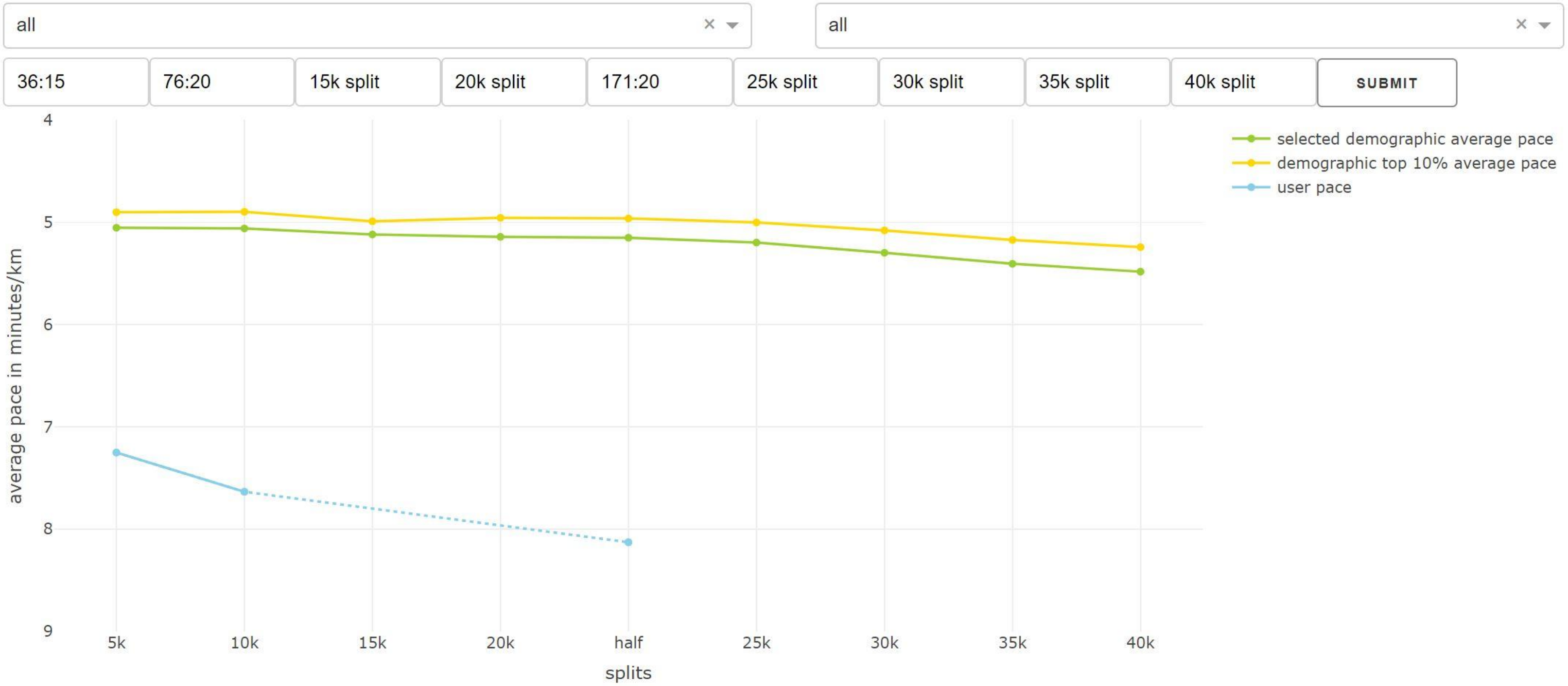
F

30-34

29:42 | 61:12 | 15k split | 20k split | half split | 25k split | 30k split | 35k split | 240:00

SUBMIT
average pace in minutes/km
4
5
6
7
8
9
5k
10k
15k
20k
half
25k
30k
35k
40k
splits
selected demographic average pace
demographic top 10% average pace
user pace

Example use case 1: Novice runner



all

all

36:15

76:20

15k split

20k split

171:20

25k split

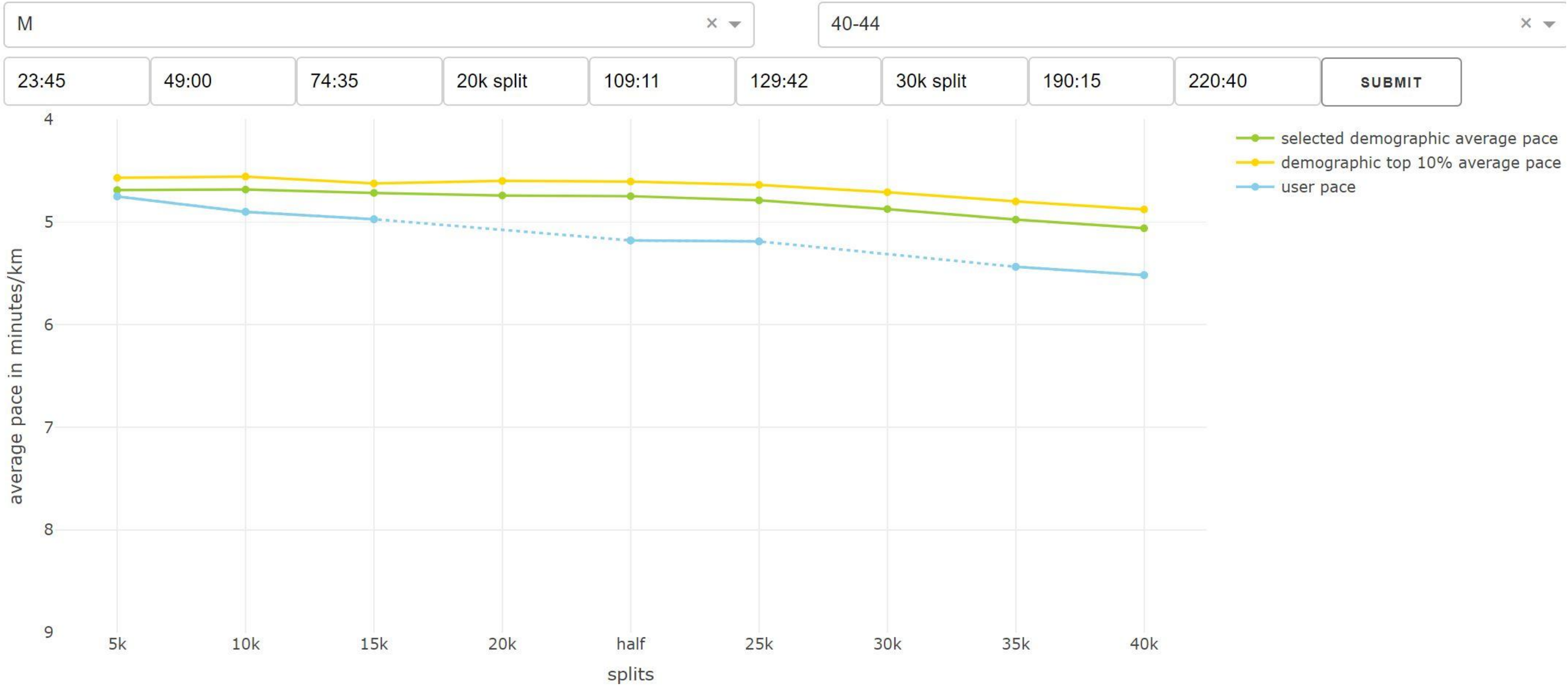
30k split

35k split

40k split

×
SUBMIT
selected demographic average pace
demographic top 10% average pace
user pace
average pace in minutes/km
4
5
6
7
8
9
5k
10k
15k
20k
half
25k
30k
35k
40k
splits

Example use case 2: Trained runner



M

40-44

23:45

49:00

74:35

20k split

109:11

129:42

30k split

190:15

220:40

× ▾
× ▾
SUBMIT
selected demographic average pace
demographic top 10% average pace
user pace
average pace in minutes/km
4
5
6
7
8
9
5k
10k
15k
20k
half
25k
30k
35k
40k
splits

Struggles

- No/different split values for years before 2013 for most marathons limiting the data scope
- Age bucket challenge- deriving age based ranks and buckets for uniformity across datasets
- London age bucket challenge- lowest bucket with range 18-38 which is too wide
- The data does not include runners who did not finish, so we don't know what went wrong with them
- Decisions about which splits to require users to produce predictions like estimated rank or split ratio; there is a trade-off between usefulness of output and accessibility of the tool

Work in pipeline

- Finish data acquisition
- Finalize data merging and formatting
- Finalize clear informative and instructional text within app
- Add remaining features if possible (rank projection, fatigue zone)
- Add unit tests
- Better organize code
- Finalize package framework

Future Work

- Update data year after year for given marathons
- Include more major marathons globally
- Develop a high intensity training version for trained marathon runners
- Suggest training plans based on the routine and data stored
- Include weather data and some features like elevation gain to be used in predictive modeling for probable ranks
- Real-time user comparison with runners (users) using the tool globally

Lessons Learned

- Setting up consistent environments from the beginning (scraping issues)
- Making basic design decisions early
- Discerning which additions are easy and can be integrated smoothly towards late in a project (new data in consistent format) and which are too big (new interactive features)