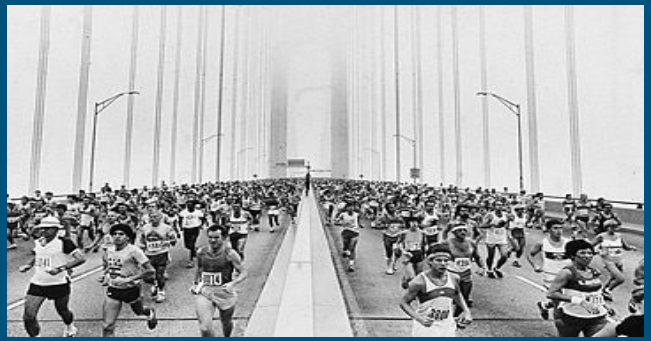
Dashathon

Marathon Training Tool for Runners

Marathons: a brief history

The Abbott World Marathon Majors is a championship-style competition for marathon runners, currently comprises of annual races for the cities of Tokyo, Boston, London, Berlin, Chicago and New York City.

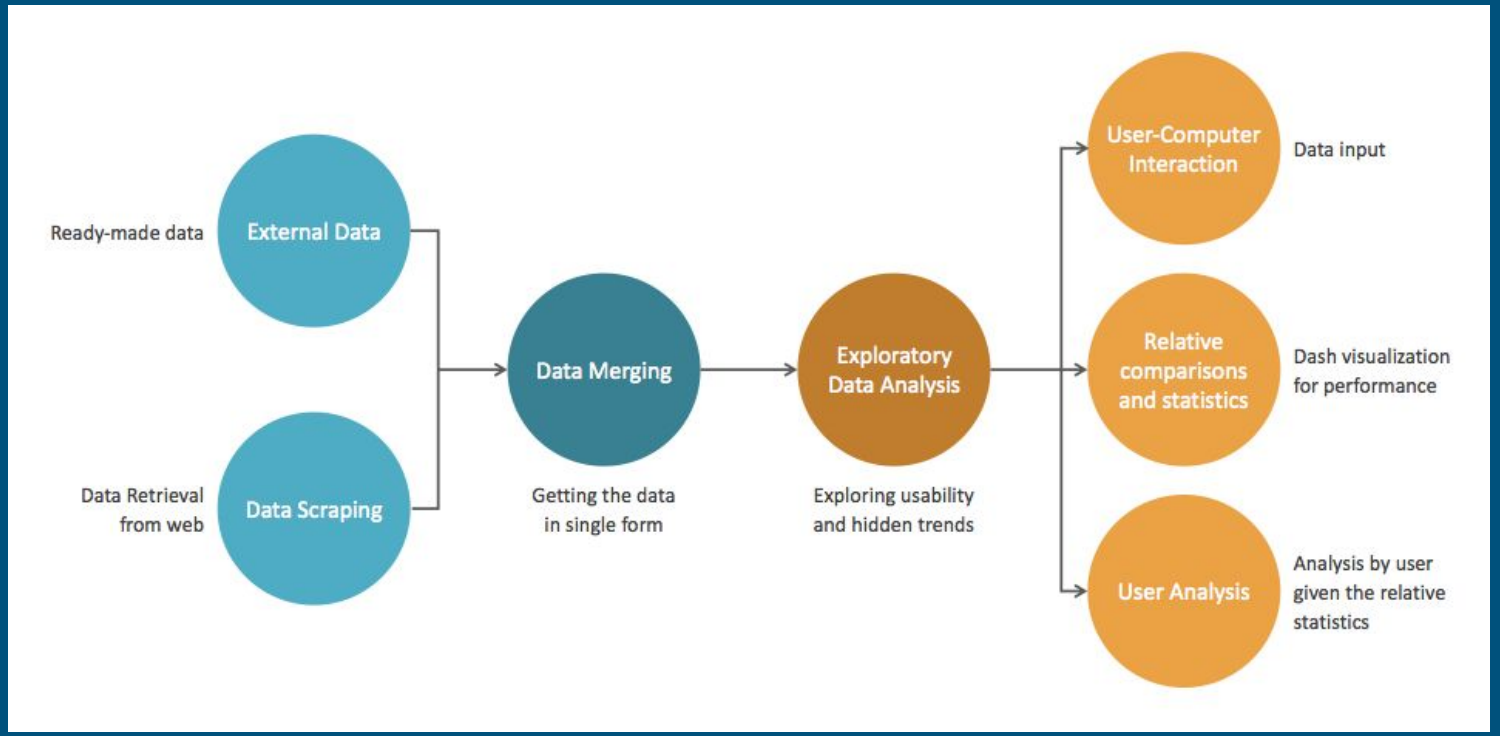
- Boston: world's oldest marathon and best-known road racing events, it began in 1897. There were 30,088 participants in 2018
- Chicago: Started in 1977, the Chicago Marathon is the fourth-largest race by number of finishers with 44,571 in 2018
- Berlin: Initiated in 1974, it is one of four world-wide 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



Motivation: our primary market

- Number of serious runners has skyrocketed with nearly 20 million runners who have entered and finished an organized event from 2008-18.
- The worldwide growth in participation from 2008 to 2018 was 49.43%
- North America has seen a growth of 21% while the primary market, Europe, grew by 43% from 2008
- Of the 443,878 marathon results recorded for the USA, an amazing 196,586 were women.

Design: component analysis



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.

- 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.

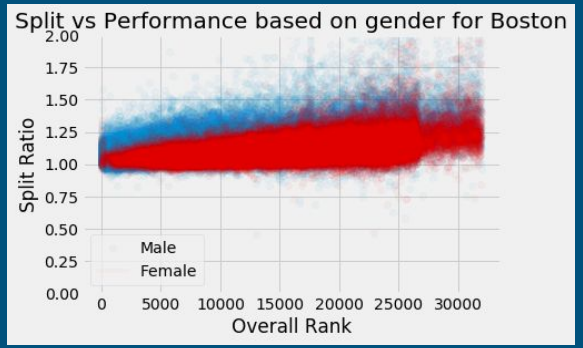
- Bib number for the year is the unique identifier of the runner
- The splits that we have are: 5k, 10k, 15k, 20k, half-way, 25k, 30k, 35k, and 40k
- Age on race day and the gender, along with the age bucket of the given marathon
- Runner's overall pace
- Runner's official finishing time
- Runner's overall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

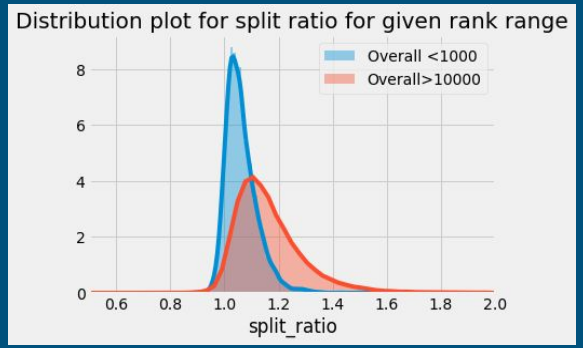
What do these features mean?

- 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

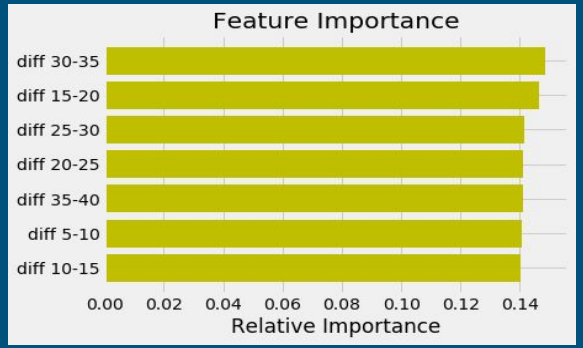
Exploratory Data Analysis



- 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?

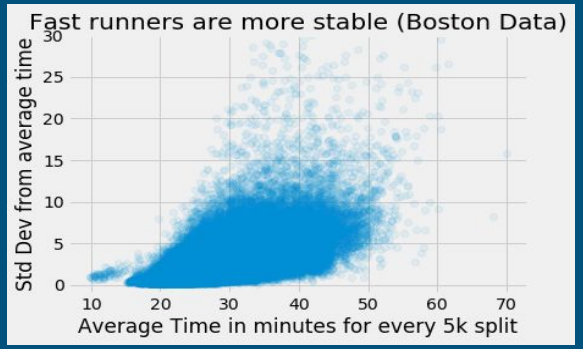


- The better runners have a less positive split than worse runners
- Runners with greater ranks tend to run a very positive split

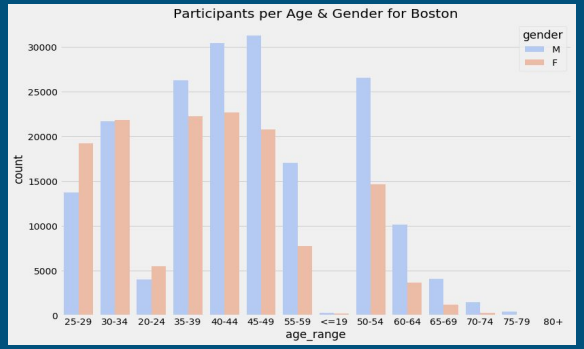


- A simple random forest regression to understand which splits contribute most to the final completion time
- All of the splits contribute almost equally. Is constant pace another popular technique?

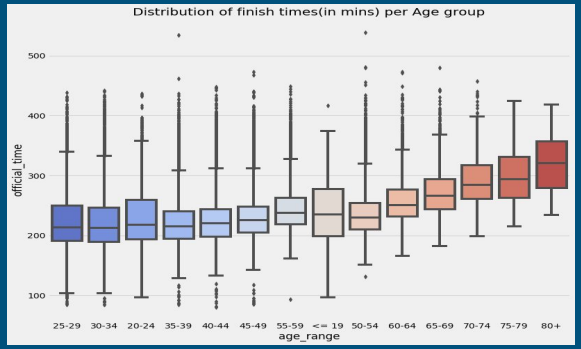
Exploratory Data Analysis



- 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 a little younger than men
- Age 40-44 has highest overall finishers



- As the age group increases, we see a slight dip in the official time to cover the race (lowest at 30-34)
- As the age increases, beyond a certain point, the official time increases

Merging the Datasets

Our current data comes from multiple distinct sources. Here is how we progressed:

- 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.

It is time for a demo. Let us look into the following use cases:

User Profile: novice runner

User is more distance focused. Runner keeps track of distance covered in given time to achieve basic distance milestones (like 5k,10k,15k, etc. splits) during that run.

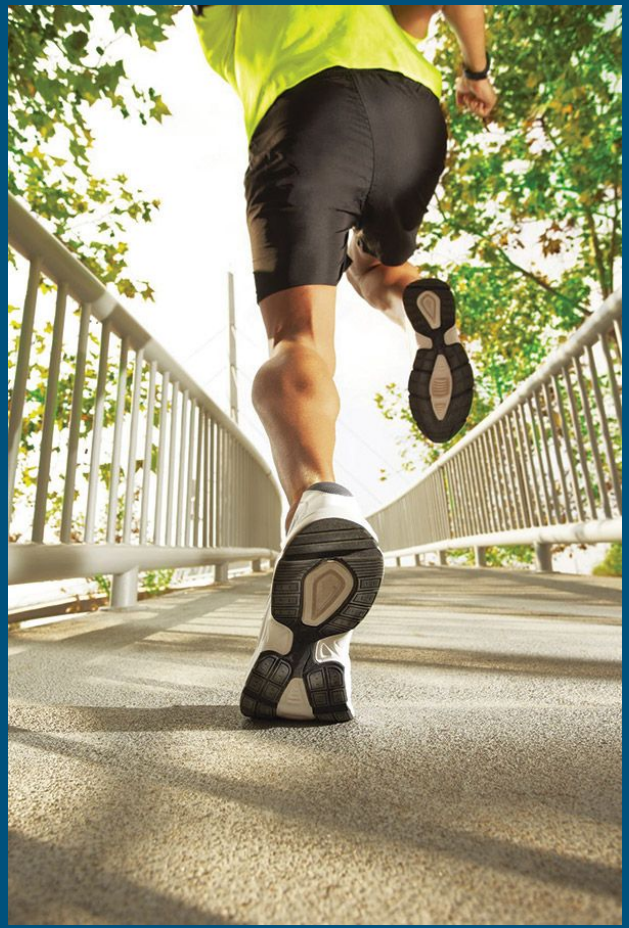
- Once the distance milestone has been selected by the user in the tool, and the age and gender has been selected, the user is able to see their performance in their demography for the chosen milestone relative to the past runners.
- The tool also enables the user to know where the user's fatigue zone during the run.
- This user is more interested in knowing where he lies with respect to the average performance of a runner in the marathon.
- They also want to understand whether their split ratio strategy was improving and how their pace was.

Use Case 1 Demo

User Profile: trained runner

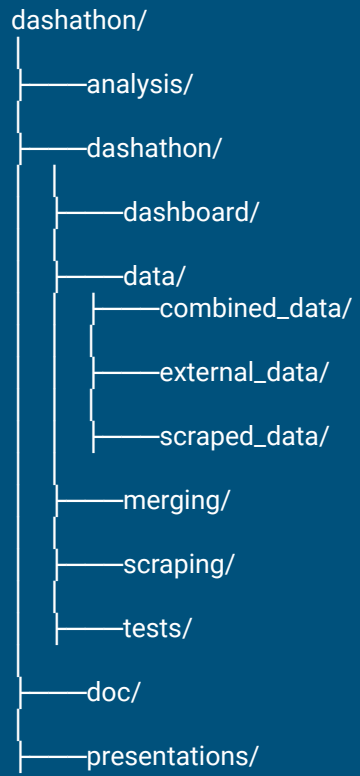
User is more time and strategy focused. User is a seasoned marathon runner who has just finished a 40k run.

- The user tracks the time it took to complete the milestones enroute and enters values in the dashboard.
- The tool displays how the user’s current ranking would vary across milestones given the user’s current run in their demography (age, gender).
- It gives them information about their split ratio standing and the area of fatigue they need to work on.
- They are more interested in comparison with a more competitive crowd, like the top 10 percentile and want to observe whether their pace was constant or if they were following their split ratio strategy well.



Use Case 2 Demo

Project Structure: git repository



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
- Reading pipe delimited CSV files with commas in columns like city.

Lessons Learned

- Setting up consistent environments from the beginning (scraping issues)
- Making basic design decisions early
- The alterations in the final goal with respect to data availability
- 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)
- Inquisitiveness vs Utility: struggle with usability when compared to richness of data
- Time and task management for effective results while working in a group

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

Finish Line Crossed. Thank You!