

NYC Marathon (an2886, ap3635, al3273, ah3617)

1. Introduction

Millions of people run marathons world-wide each year. Marathons help raise money for charities, connect runners around the world, and help people to exercise and lead healthy lives while inspiring others to do the same. Our group chose to analyze the NYC Marathon Data because this marathon is a unique event which brings together extraordinary, driven people from all over the globe, from all age groups, at different levels of physical ability to push their bodies to the limit. **We were curious to find out who these people were, where they came from but also how these variables influence their performance.** Additionally, **we were interested in understanding how large is the participation of woman relative to men and what is the trend.** Finally, **we explored how to summarize the whole race by examining the performance of runners relative to each other.** The team consisted of Arthur Herbout, Antonia Lovjer, Andrea Navarrete and Andres Potapczynski.

Andrea took on the challenge of scraping the data from the NYC Road Runners website, and cleaning it using an R script which she shared with the rest of the group. The five data sets consisted of the NYC marathon data on race finishers for years 2015 to 2018. Andrea and Antonia continued to do a demographic analysis of the data, working towards understanding who were the people participating in the race. Arthur wanted to understand how the race was organized, and used variables such as “official_time” and “gender” to understand the meaning behind the “waves” and how they correlated with performance. Andres took a more in-depth approach to the data by clustering the runners according to their performance in order to analyze patterns amongst the different groups.

As a team, we worked on brainstorming ideas for the project, organizing the analysis, completing the exploratory data analysis and visualization, and creating the presentation. Design of the interactive Shiny app was completed as a group, with the development led by Andrea, and contributions by Arthur, Antonia and Andres.

2. Description of data

To answer research questions mentioned in the introduction we generated the data set by web scrapping the official site of the race (for all the details of this process look at section 2.2 below). It is worth mentioning that our data is made from only the competitors that finished the race. In 2018, our data set contains 52,669 runners. In 2017, we had 49,238 runners. In 2016, we had 50,486 runners and finally in 2015, we had 48,742 runners. This section is divided as follows. First we mention the relevant variables that we included in the data (as well as their actual R code names). Then we dive into the source of the information. Finally, we explain (or rather summarize) the intricate details of web scrapping the data.

2.1 Relevant variables for the analysis

Following is a list of the main variables used for the analysis (we omit adding the details for variables that we did not use in any of our analysis).

- **gender:** the gender of the athlete encoded as “M” for male and “F” for female.
- **age:** the age of the athlete.
- **official_time:** the time it took the athlete to finish the race.
- **gun_time:** the actual time when the athlete got to the finish line (remember that some athlete started at different times).
- **name:** the name of the athlete (we used this to keep track of how people improve or worsen their performance if they appear in other years).

- **city**: the city where the athlete comes from.
- **state**: the parse state abbreviation (only for the US) where the athlete comes from.
- **stat_name**: as above but the complete name of the state.
- **country**: the country where the athlete comes from.
- **lat**: the latitude given by the country where the athlete comes from.
- **long**: similar to above but now the longitude.
- **split_<x>k** the time when the athlete got to the x th split where $x \in \{5, 10, 15, 20, 25, 30, 35, 40\}$.
- **type**: the inferred group category. “R” is for runners, “H” for hand cyclists and “W” for wheelchair.
- **team**: the name of the team if the athlete belongs to one. For example, professional athletes might be sponsored say by *NIKE*, other athletes represent charities associations like *Team for Kids* and finally some others might belong to an amateur group like *North Brooklyn Runners*. Nonetheless the majority does not report belonging to a group and are added as *NA*.

2.2 Source

Our data set comes from the *TCS New York City Marathon Results* which is hosted by the New York City Road Runners. We scrapped the data from their official website link. Accessing the previous link renders the following view:

The screenshot shows a webpage for NYRR Race Results. At the top, there is a navigation bar with a menu icon, the title "NYRR RACE RESULTS", and a search icon. Below the title, the runner's name "Lelisa Desisa" is displayed, along with her details: "M28 | Addis Ababa | NIKE" and the period "Active: 2011 - 2018". There are three tabs at the top: "OVERVIEW", "RACES", and "COMPARE", with "OVERVIEW" being the active tab. To the right of the tabs is a link "REQUEST CORRECTIONS". Below the tabs, there is a back button "< Back", social sharing options (Photos, Share, Facebook, Twitter, Email), and race information: "TCS New York City Marathon 2018 MARATHON | NOV 4, 2018". The main content area shows race results for "AGE/GENDER M28 BIB 3 TEAM NIKE". Key results include: OFFICIAL TIME **2:05:59**, PACE PER MILE **04:49**; PLACE OVERALL **1** of 52,705, PLACE GENDER **1** of 30,582, PLACE AGE-GROUP **1** of 2,876; PLACE AGE-GRADED **1** of 30,582, TIME AGE-GRADED **2:05:59**, PERCENTILE AGE-GRADED **97.6%**; GUN TIME **2:05:59**, GUN PLACE **1**. A "Splits" section lists times for 5K, 10K, 15K, 20K, and HALF distances. Navigation links "LIST VIEW" and "GRAPH VIEW" are also present.

SPLIT	TIME
5K	0:15:43
10K	0:30:51
15K	0:45:48
20K	1:00:44
HALF	1:03:57

Figure 1: View of the webpage that we scrapped

Thus, as it can be seen above the information that the website provides falls into two categories.

- **Demographics:** Age, gender and country / city
- **Performance metrics:** Official time, pace per mile and the time per 5 kilometer split

The main difficulty with this data is that it is embedded in a web page. Even though it is easily accessible it is hard to download locally! Thus, we had to figure out how to web scrape it.

2.3 Web Scrapping

Web scrapping was more time consuming than we thought, since the website was built under javascript in order to scrape we needed the use of `selenium` which is an is a WebDriver tool for automation that basically connects to a driver and retrieves the rendered html. From the html we were able to obtain the data using a really well-develop package of `BeautifulSoup` but it still required us to overcome two main obstacles (1) understanding the web scrapping process and (2) to come up with the template that our program should use in order to find the information.

To get a complete understanding of the process, we made intensive use of different resources online: either YouTube videos or other Q&A websites such as `StackOverflow`. A particular source that we found useful was this article ([link](#)). Moreover, we became acquainted with the myriad of details that were not evident when we started the process. First, we had to distribute our work in a computer engine on the cloud. The difficulty is that opening a web driver each time to obtain the render html is a really slow process, specially that we needed for each finisher. For example, we have to run a long `for-loop` where we have to make a different connection for each of the over 50 K participants of the marathon for each year (this times the number of year that we downloaded). Additionally, we had to set-up the computer and parallelize the process and let it do the work (which took approximately 3 days)

In term of coming up with a template for our program to run. We had to learn (actually before starting D3) the structure of the html website and embed that knowledge into a script for `BeautifulSoup` to perform its magic. Furthermore, this is a sensitive procedure. Altering any location of an element in the web page renders the script useless. Thus, on the one hand, we had to make several “robustness checks” for our script before letting it run (because we risk losing days of work). The good thing is that the same template was able to be used for all participants across the four years of data we scrapped.

At the end, we were quite happy that we were able to download the data. Mostly because we were doing an analysis that excited us but also because, due to the difficulty of the process, we knew that not many people had done this analysis before.

3. Analysis of data quality

Due to the web scrapping procedure we were afraid that the data would have a bunch of errors, but actually it did not. We divided our analysis for data quality into four sections. First, we analyze the general missing values patterns. Second, we delve into the one source of missing values: *messy locations*. Then, we analyze the other source of missing values: *inconsistent time formats*. Finally, we comment on the nature of outliers in this data.

3.1 Messy Locations

Sometimes the geographical information displayed in the webpage related to the city of the runner, sometimes to her state and some others times to the county she was born (definitely our web scrapping program was not so smart as to make those types of swaps). Then we had to make a decision on what to do about this: either to spend quite some time on mapping manually all the cases or do some kind of parsing and leave aside the cases without a match.

A priori we were unsure what to do, we had mixed feelings about not using this information, which was a key variable for our analysis. Hence to make a decision first we assessed how bad our problem was. For this we ran our parsing procedure to see how many cases lack a match. Our parsing procedure consisted on two steps: first, we would try to parse the states and for the remaining we would pass our country mapping (which was tractable). Fortunately, we got quite lucky and the procedure almost mapped completely off-the-shelf. The reason is that the majority of the inputs that had a state name in the geographical location where from the USA and also because the R function `state.abb` was quite robust and it allowed us to get the state abbreviations. Therefore, we left the rest of the cases as NA and move on into adding the latitude and longitude of the place.

3.2 Inconsistent Time Formats

We have never had a good experience when working with a time series data sets and this was, unfortunately, not the exception. We made extensive use of the available tools to go from formats like “01:00:23” to an actual meaningful time stamp (on this, it appears as if rather than spreading the best practices of presenting time variables in a data set, the world is rather creating more sophisticated tools to go over every eventuality).

The previous discussed problem was multiplied by all the columns that contained time data which was quite a few: the finished time, the time per 5 k split, the official time and many more. Thus after running our parsing function on each column we filtered out any observation that had at least one missing value (that is, not a parsable time) in any of the main variables such as `gun_time`, `official_time`, and each of the splits `split_<x>k`.

3.3 Outliers

Outliers in this context come either from poor performing runners, excellent performing runners and other type of competitions categories (wheelchair and hand cyclists). Thus, the notion of *outlier* here stems from mixing many different cohorts of people (professional athletes, standard competitors, wheelchair and hand cyclists) in the data rather than by having some plausible error in a certain recording of a variable. Although there were some people that finished the marathon in around 7 hours. But we are uncertain if this is a mistake or maybe they burnt out and finished the race as they could.

Hence, after separating the data into different groups, now there where not evident outliers (in the sense of points going past the whiskers in the box plots). For example, the top 100 runners all stayed very tied. Whereas the latter group had similar running times but there was more variance.

4. Main analysis

This section contains the main results of our exploratory data analysis. Each of the subsections below addresses one of the research questions that we posed in the introduction.

4.1 Demographic

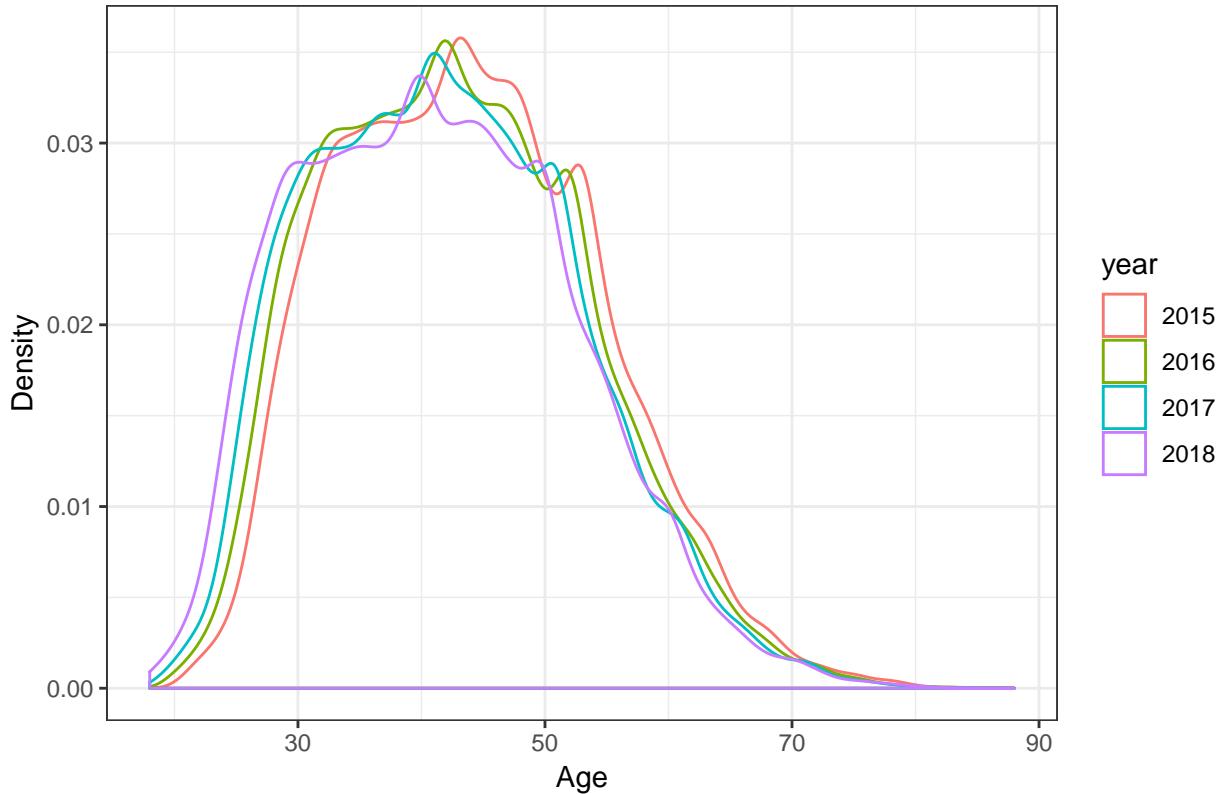
4.1.1 Age Distributions

Performance in athletics is often impacted by the age of the athletes. For example, most of the participants in the Olympics and in international sports competitions tend to be people on the younger end of the spectrum. Since marathon running is a test of endurance, it would be interesting to see if similar trends exist amongst the participants of the NYC Marathon. We can plot density curves of the age of the finishers in each year for which we have data.

From a first glance, we notice that the distribution of the data appears to have an approximately normal shape, with some flattening out at the top of the curve, and a clear mode. The youngest participant is 18, since the rules of the race most likely forbid minors from participating, and the oldest is 87.

Overall, there are more younger people finishing the race in comparison to older people, with a large portion of the participants between the age of 30 and 50, and this trend remains fairly constant across the years. The median age is between 43 in 2015 and 40 in 2018, and there is a slight trend towards the participants becoming younger. This is evidenced by the small leftward shifts in the curves from one year to the next.

Age Distributions from 2015–2018

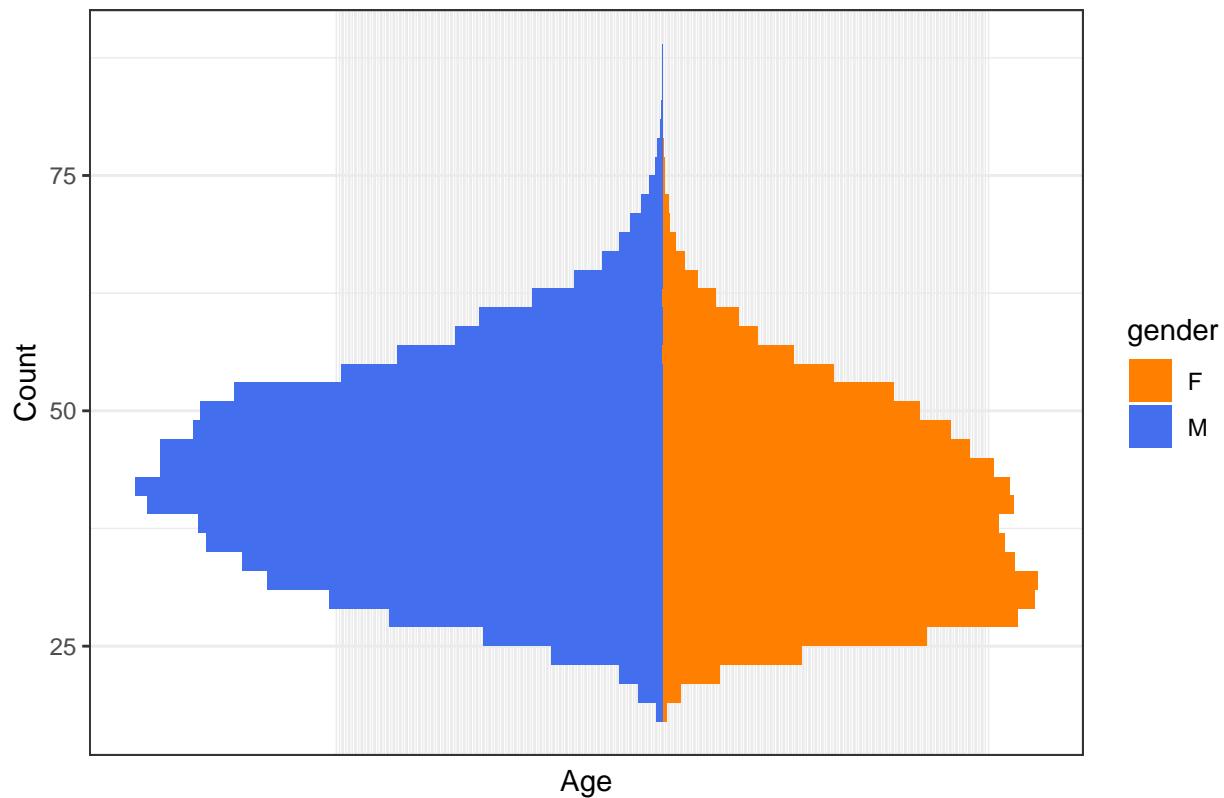


4.1.2 Age distribution by gender

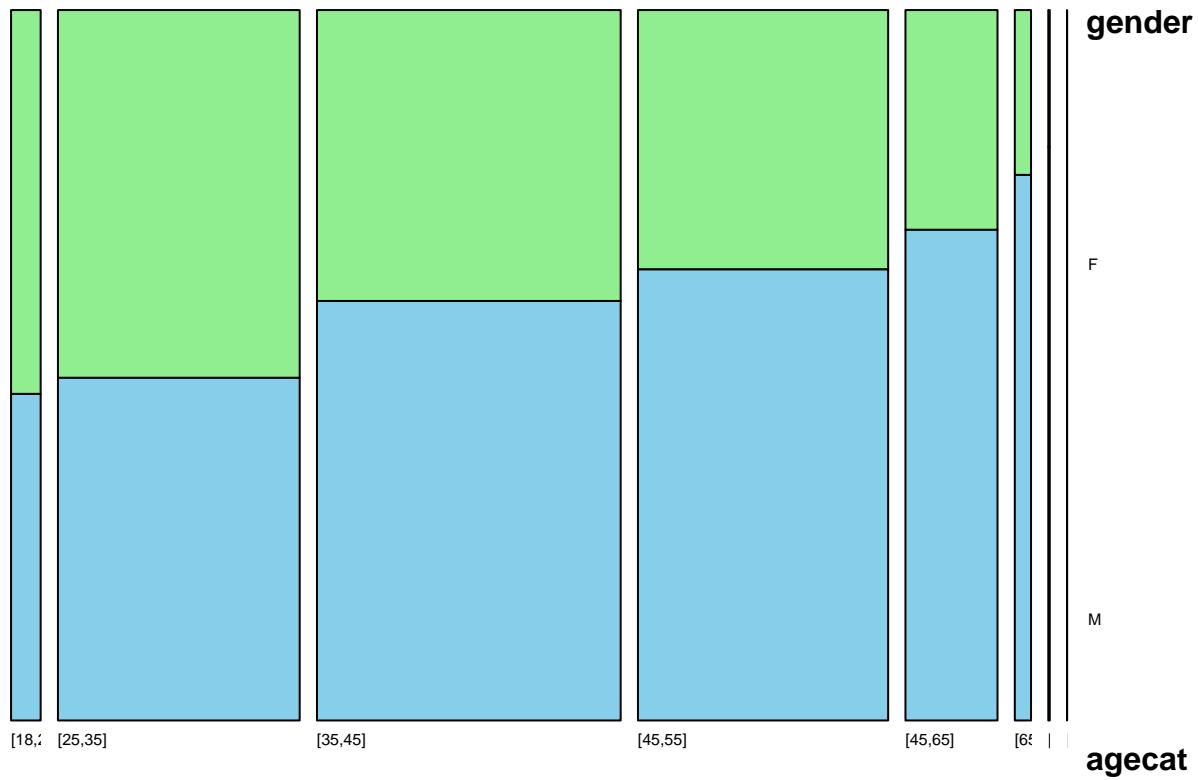
The next step would be to examine the age distributions across the different genders. The data collection process only allowed for the representation of two genders: Male and Female, and so we will restrict our analysis to these two genders.

It is interesting to see that the distributions seem very similar across the two genders, with unimodal, approximately normal shapes. It is interesting as well to see that there are more younger women amongst the finishers than there are younger men. Overall, the men seem to be a few years older on average in comparison to the women.

Distribution of Age By Gender



We can also examine this data in the form of a mosaic plot. It is more clear from this plot that there are more men in the race than there are women. One of the benefits of a mosaic plot is that relative proportions between different categories are easier to determine, but one of the downsides is that distributions cannot be determined.

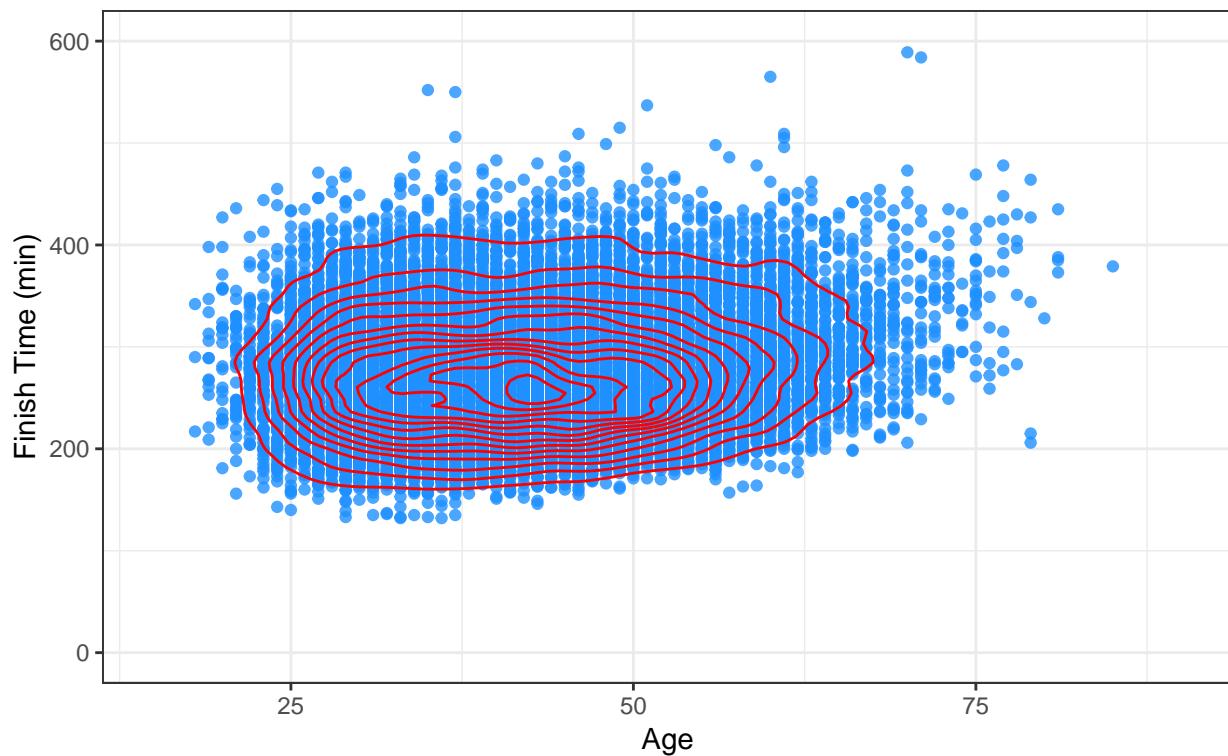


4.1.3 Finish Times by Age

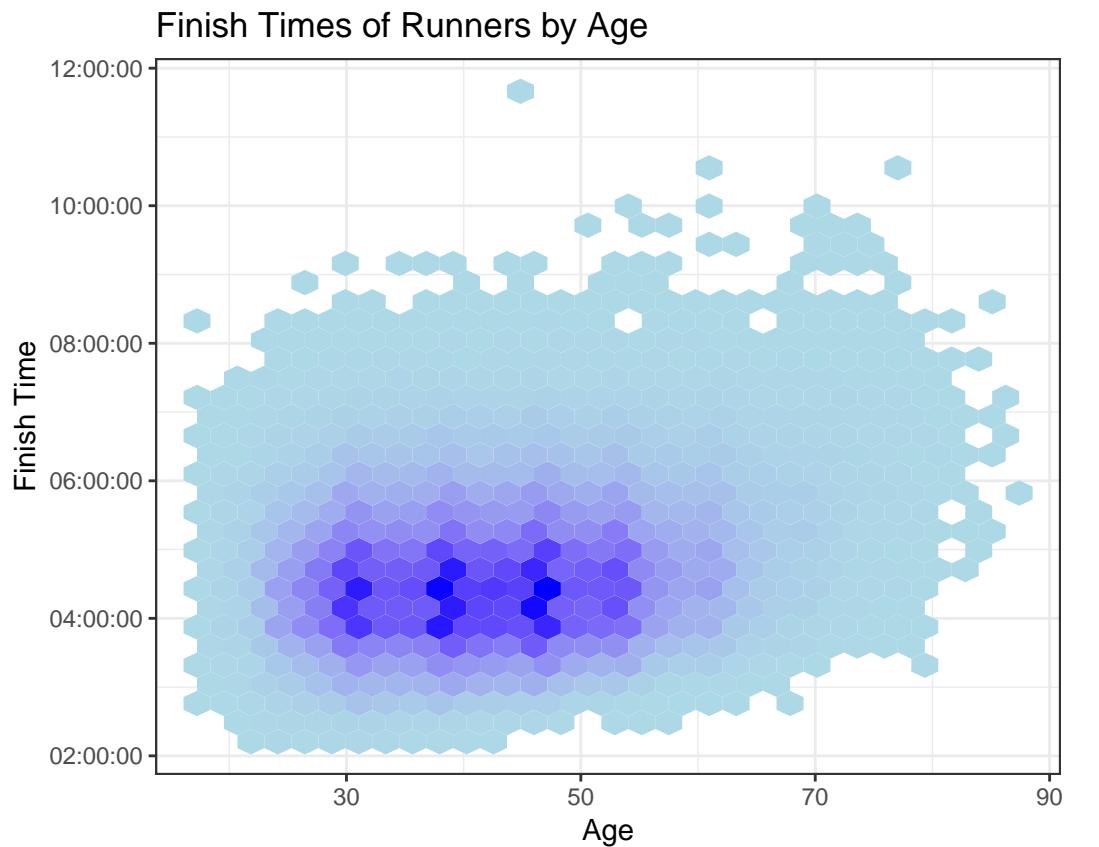
Next we examine the finish times by age of the runners. Here we will look only at the actual “runners” of the race, excluding hand cyclists and wheelchair racers, as their times would be significantly faster than the runners. We can try to display the data in several formats, the first will be a scatterplot where we take a sample of the data and plot age on the x-axis and finish time in minutes on the y-axis. We can see that there are a lot of runners aged 35 to 50 and running time between 4 and 5 hours.

Finish Times of Runners by Age

Sample Size: 20,000



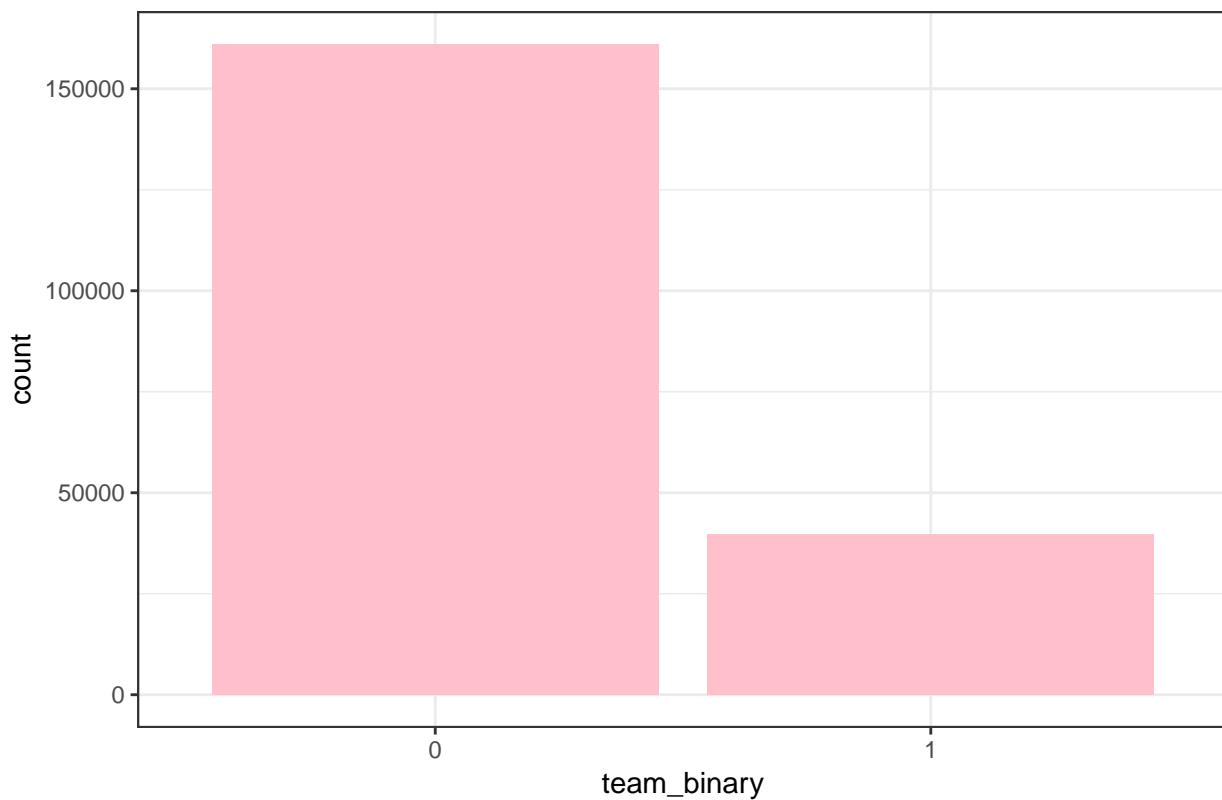
Another way to visualize this information is by creating a heat map with the same axis as in the previous plot. Here we will use the entire data set, and have the darker colors represent the value pairs where there are more data points present. We see that the highest concentration of runners can be found between the ages of 30 and 50, and between 4 and 5 hours. Another observation is that the older people get, the fewer fast runners there are. This is evident by the fact that there are no runners in the bottom right hand corner. What is also interesting is that all of the fastest runners cover a fairly wide age range from 20 to 42.



4.1.4 Analysis of Teams

All together there are 651 different teams that entered the NYC Marathon Race in the running category. Many of these teams are groups of volunteers that joined their company team or a local charity team. They are also amateur running clubs, and finally professional teams that are sponsored by large athletic equipment firms such as Nike and Adidas. If we look at all of the runners, we see that 39815 of the runners were part of a running team, while 160998 we not, which is over 80% of the runners. Therefore, it's only about 1/5 of the runner that finished the race were part of an organized team. In the graph below, 0 indicates the non-team runners, and 1 indicates the runners on teams.

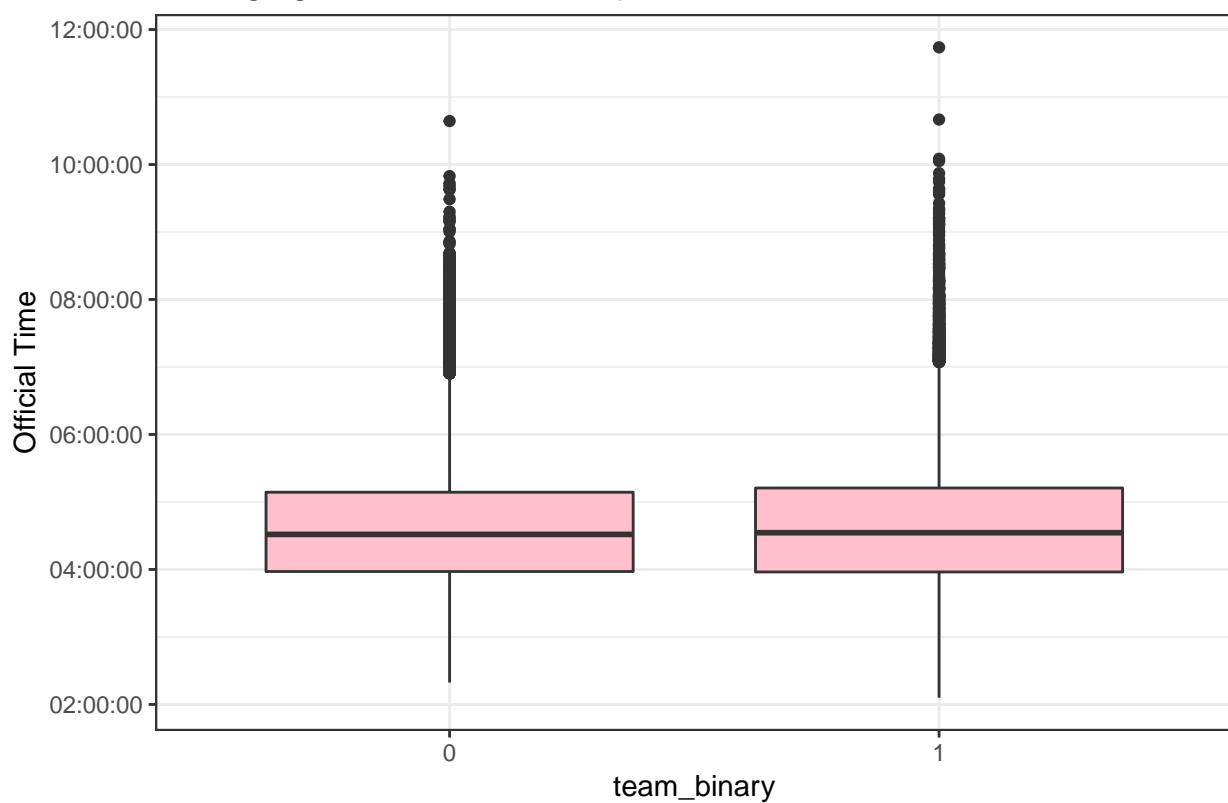
The majority of people do not belong to a team



Next we look at box plots for the official finishing times for team runners vs. non-team runners. Surprisingly the distributions look almost identical. The medians are very close to one another for the two groups, as are each of the other quantiles.

We can see that the outliers are slightly more extreme for the team runners in both directions. The fastest runner is part of a team, and the slowest runner is also part of a team.

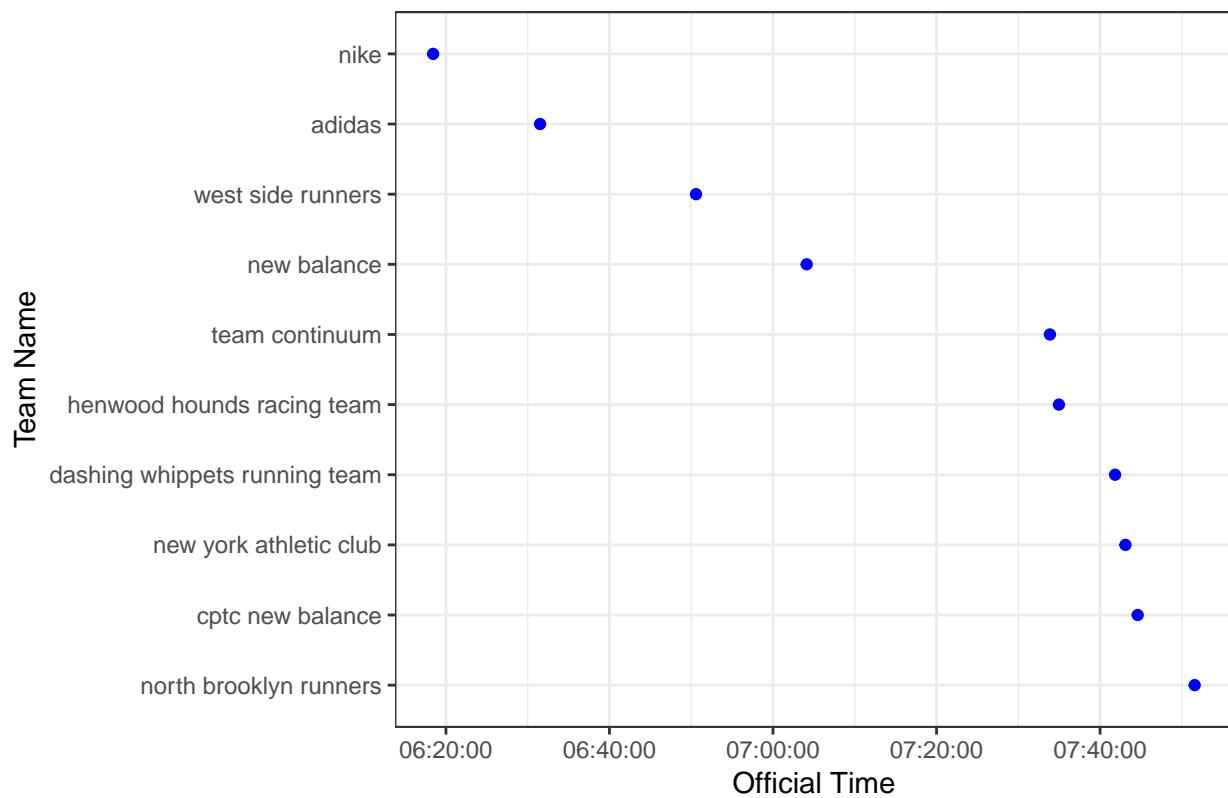
Belonging to a team does not predict better outcomes



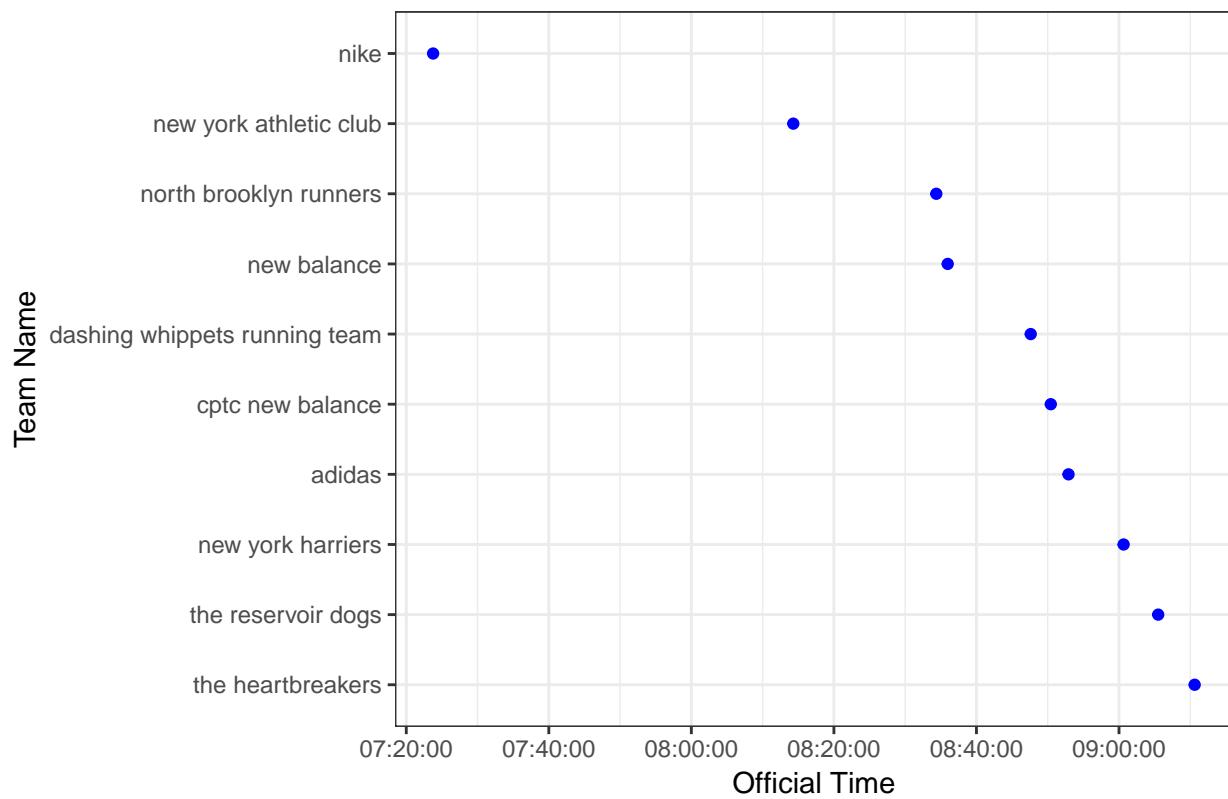
Now we can look within the teams to identify the best teams for women and for men. To do this we looked at teams that contained more than 5 members, and ranked the team members according to their performance. Then we took the sum of the top 3 runners per team, and again ranked the teams.

It is not surprising that many of the team names repeat for both genders. Additionally we can see what many of the names would be familiar to people, such as Nike, Adidas, New Balance and the New York Athletic Club

Top 10 Teams – 2018 Male



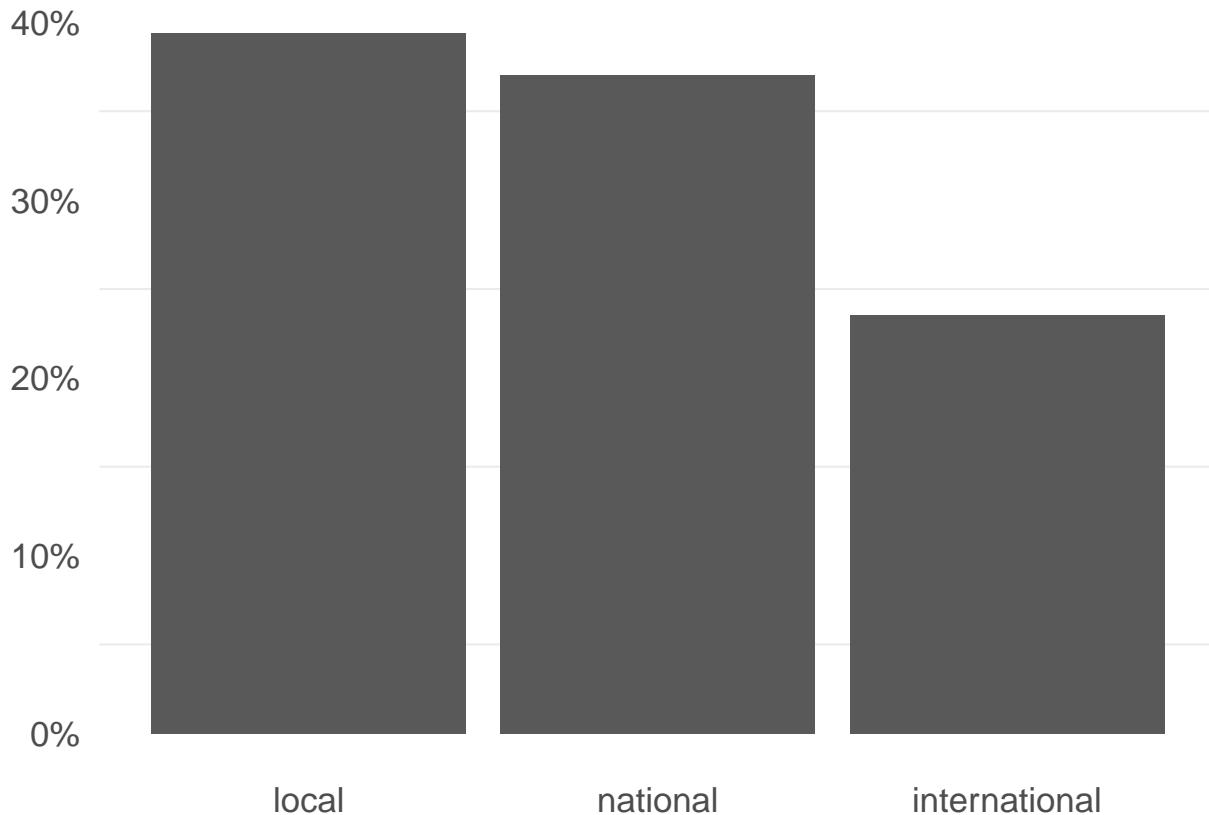
Top 10 Teams – 2018 Female



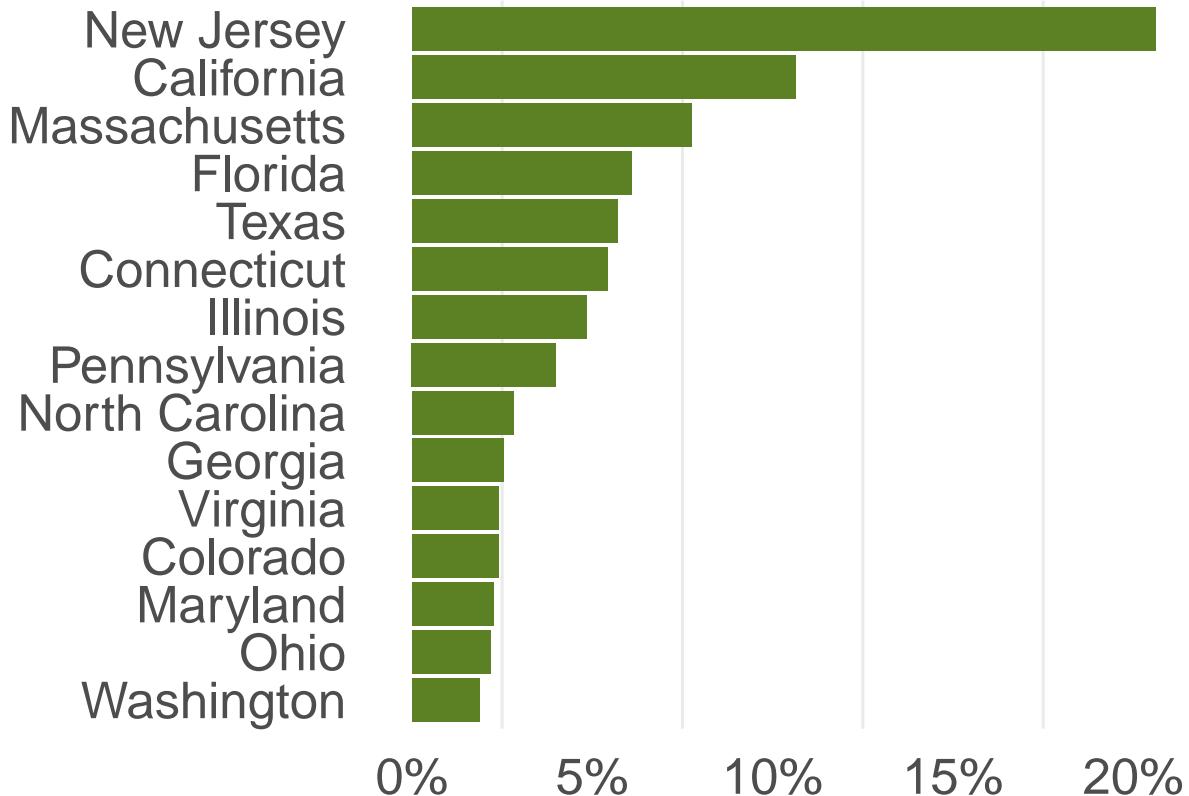
4.1.5 Location

We can group the location data into three groups: local, national, and international. From the table below, we see that the proportion of representation that these three groups receives remains fairly constant over the 4 years. Around 40% of the finishers are local, meaning that they reside in New York. There are around 37-40% representing the USA excluding those from NY. And finally, around 20% of the participants come to NY from other countries in the world. This information is summarized in the bar chart below where the data has been aggregated across years.

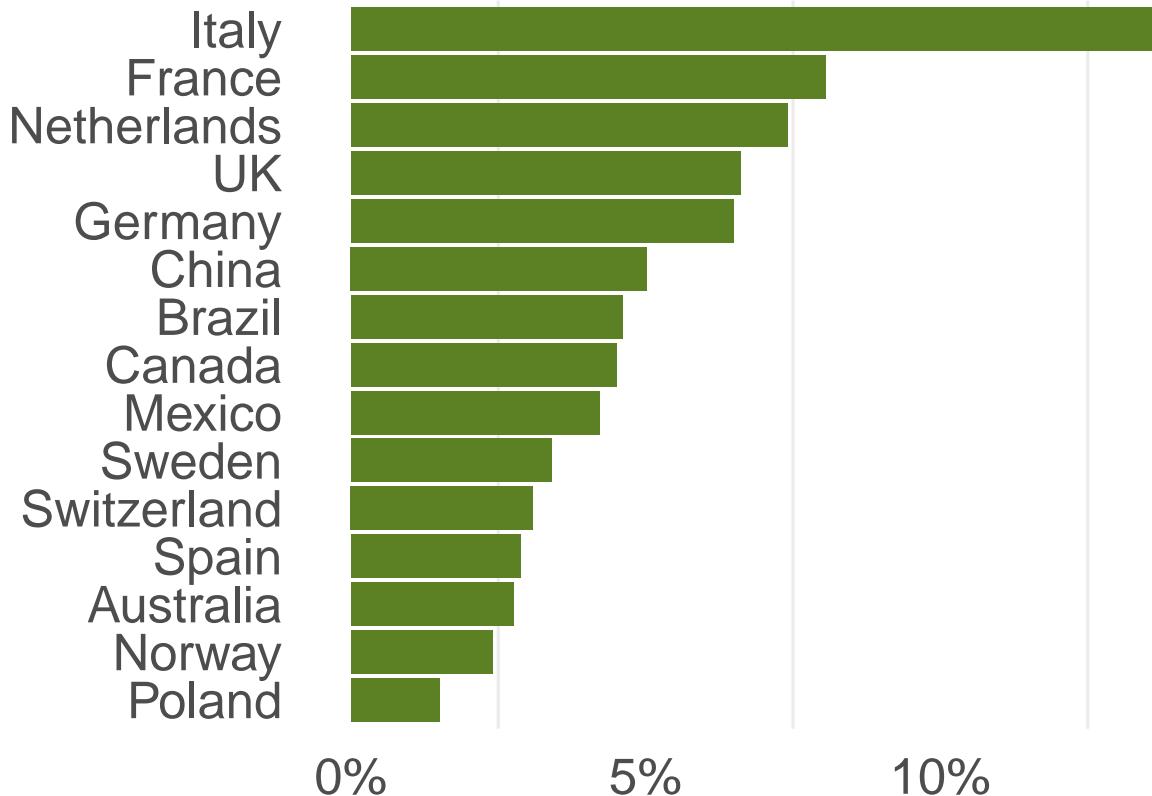
```
## # A tibble: 4 x 4
##   year   local   national   international
##   <fct> <dbl>    <dbl>        <dbl>
## 1 2015   0.392    0.374       0.232
## 2 2016   0.386    0.398       0.214
## 3 2017   0.402    0.374       0.224
## 4 2018   0.394    0.370       0.235
```



Excluding New Yorkers from the data, we can further investigate which US States are most represented amongst the marathon finishers. In the horizontal bar chart we see that after New York, New Jersey has the most participants, followed by California and Massachusetts. These results are not surprising as New Jersey and Massachusetts are very close to New York. Additionally, California is the most populous state therefore is expected that it would be heavily represented, despite its lack of proximity to New York.



We can examine a similar plot at the country level. After the US, Italy by far has the most represented country, followed by several other Western European countries, China, Brazil, Mexico, Canada and Australia. It is interesting to note that Europeans have higher rates of participation in the marathon compared to countries that are close to the US, including immediate neighbors Mexico and Canada.

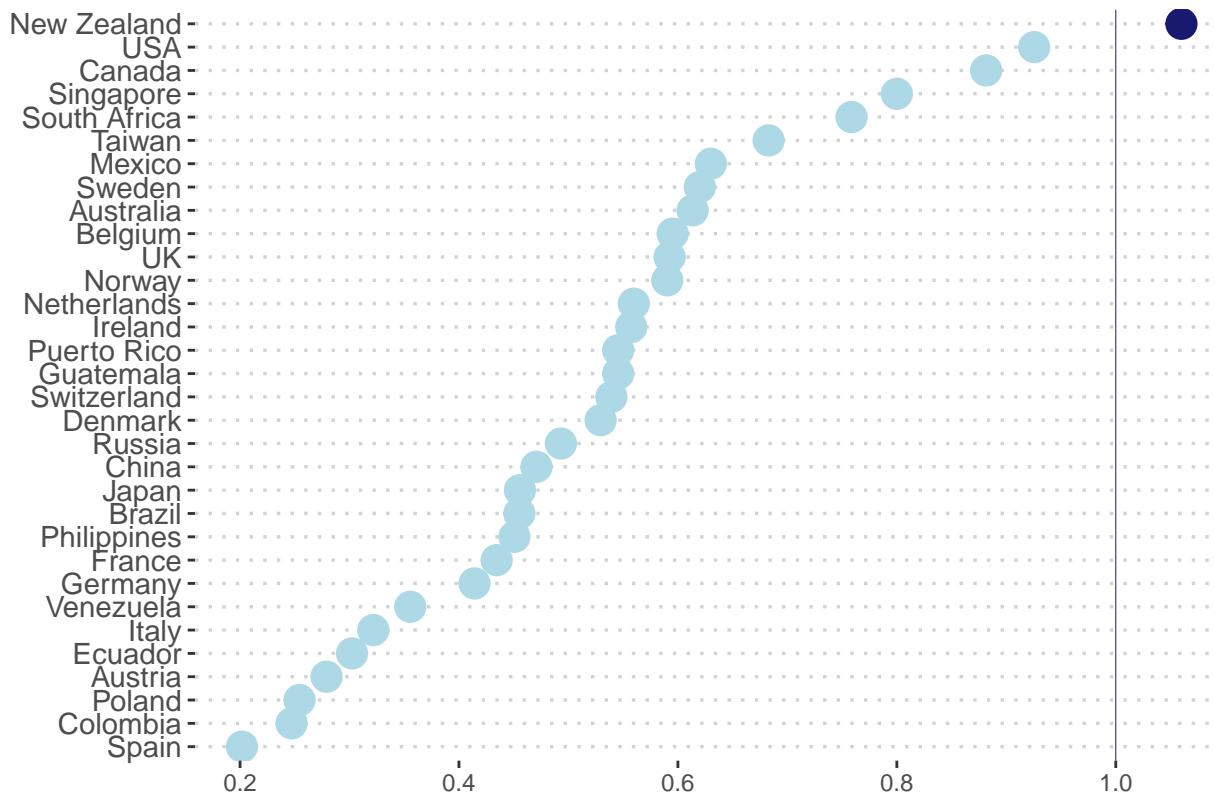


4.1.6 Women Ratio

The ratio of men and women participants in sports has long been an important topic of discussion. We can examine to what extent women are equally represented amongst the NYC marathon finishers according to their country of origin. In the cleveland dot plot below we plot the fraction of runners from a country that are women, with the country names on the y-axis ordered by decreasing proportion. The vertical line indicates equal proportion.

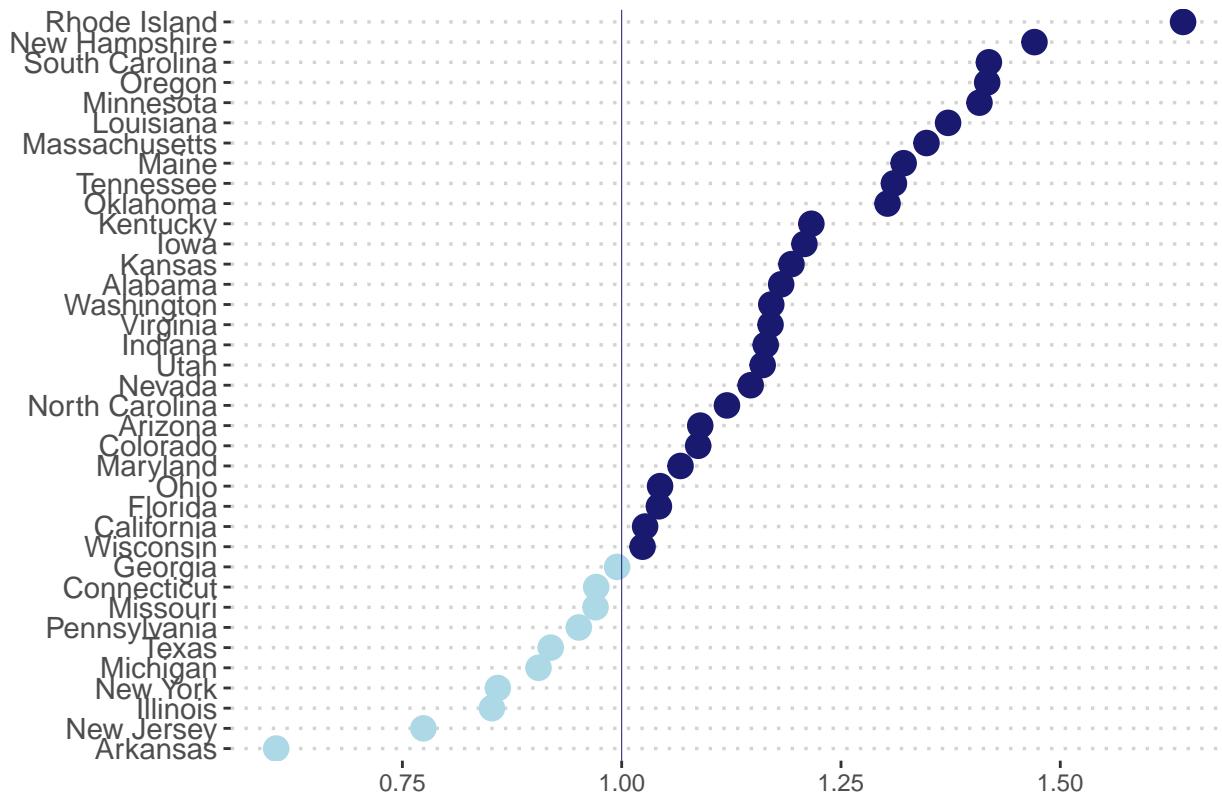
We note that in 2018, only New Zealand is represented more highly by women than by men. Russia is 19th on the list having a proportion of around 0.5, meaning that the remaining 122 countries represented had more than double the number of men compared to the number of women. From the table we see that the aggregate proportion has remained relatively stable over the last 4 years at around 0.7 for the runners. But from the visualization it is clear that many of these women are coming from only a handful of countries.

Ratio of woman and men by Country



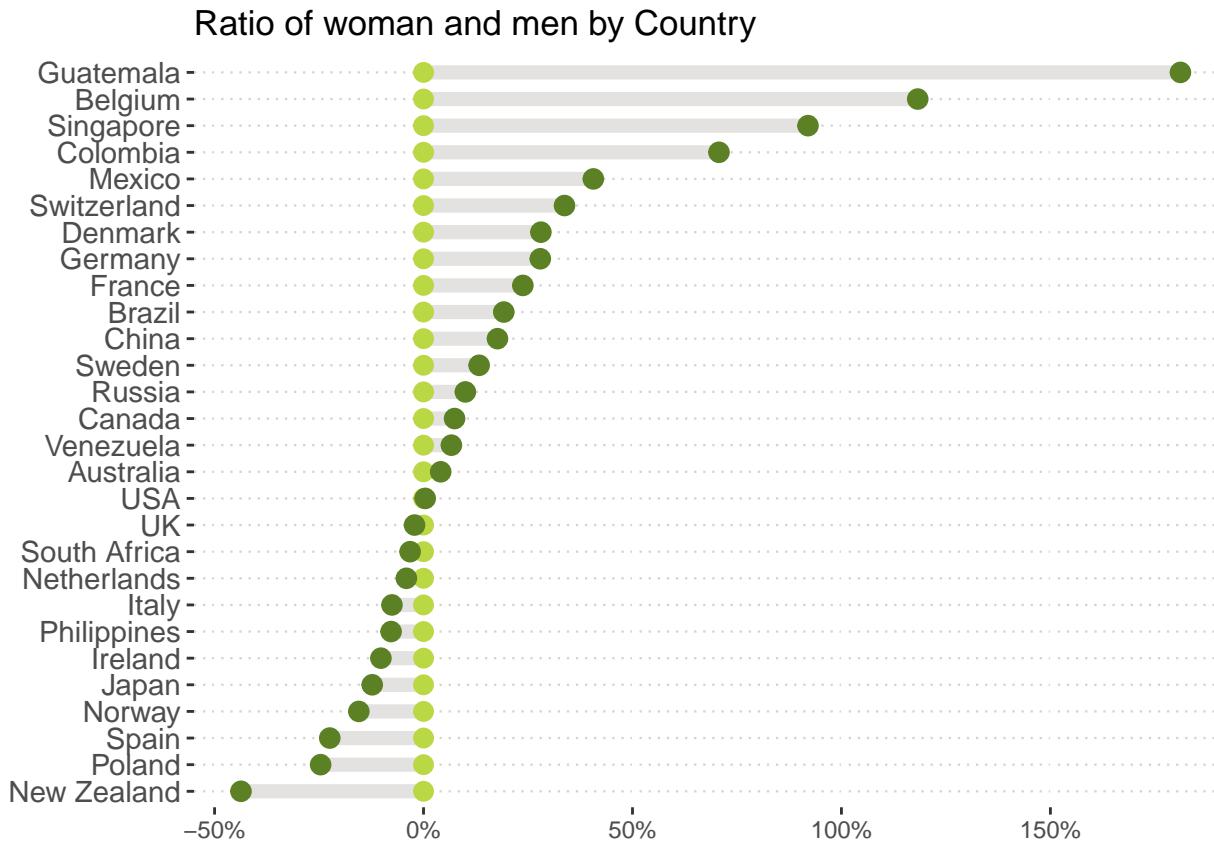
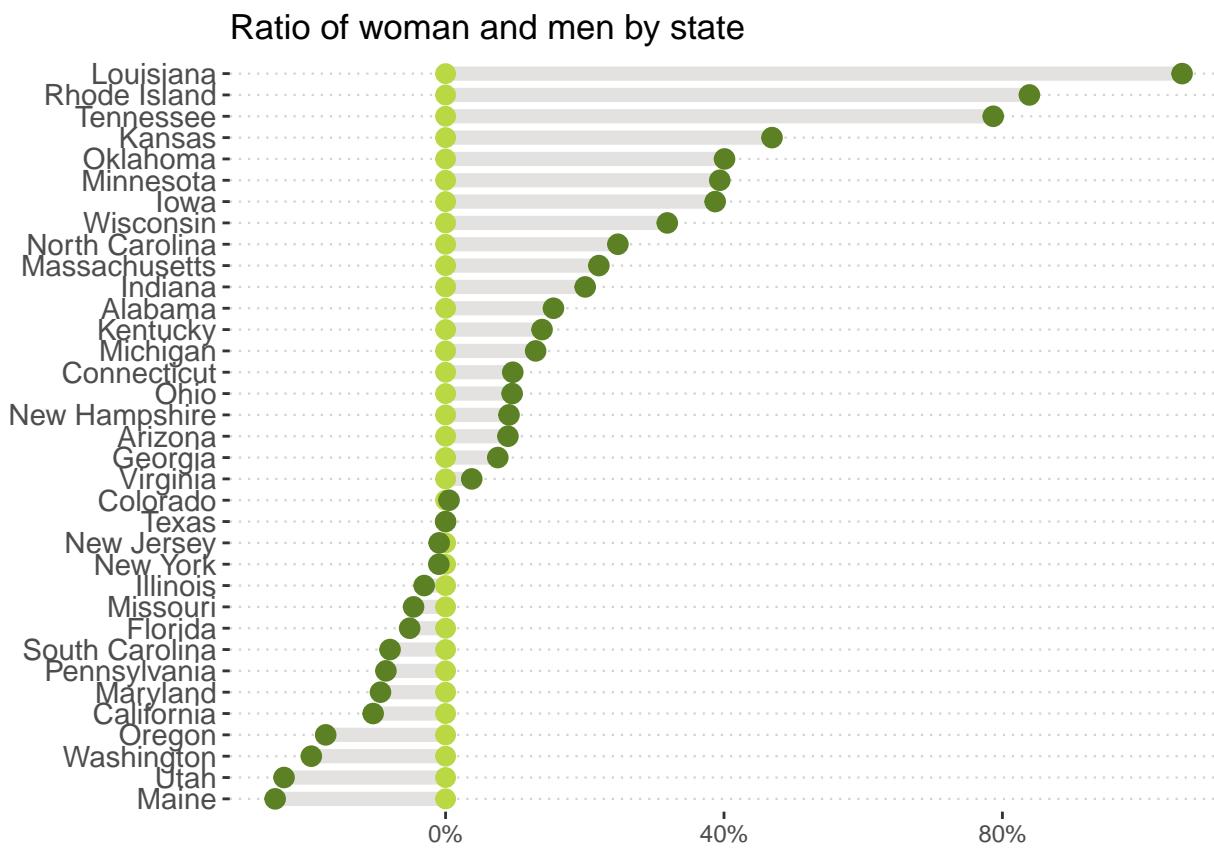
On the US state level, however, women are the majority of finishers for 27 states, including California which is the third most represented state. Many of the other states in the aforementioned 27, however, do not have many participants, and so the ratio of women to men is close to, but not equal to 1.

Ratio of woman and men by state



Visualizing the change in these ratios over time shows us that many states and countries have had a positive change in the number of women they have participating in the NYC marathon from 2015 to 2018. We can note the trend that many Southern and Mid-Western states including Louisiana, Alabama, Kentucky, Tennessee, and Kansas have increased their fraction. Similarly, several countries have increased the number the fraction of women representing their country several times over.

Despite all these positive individual country-level and state-level shifts in the gender break-down of the people representing their countries or states, the overall ratio of women to men has remained roughly constant over the past four years. Therefore, while some locations are sending more women, others are sending less, and the fraction remains constant.



4.2 Running Variables

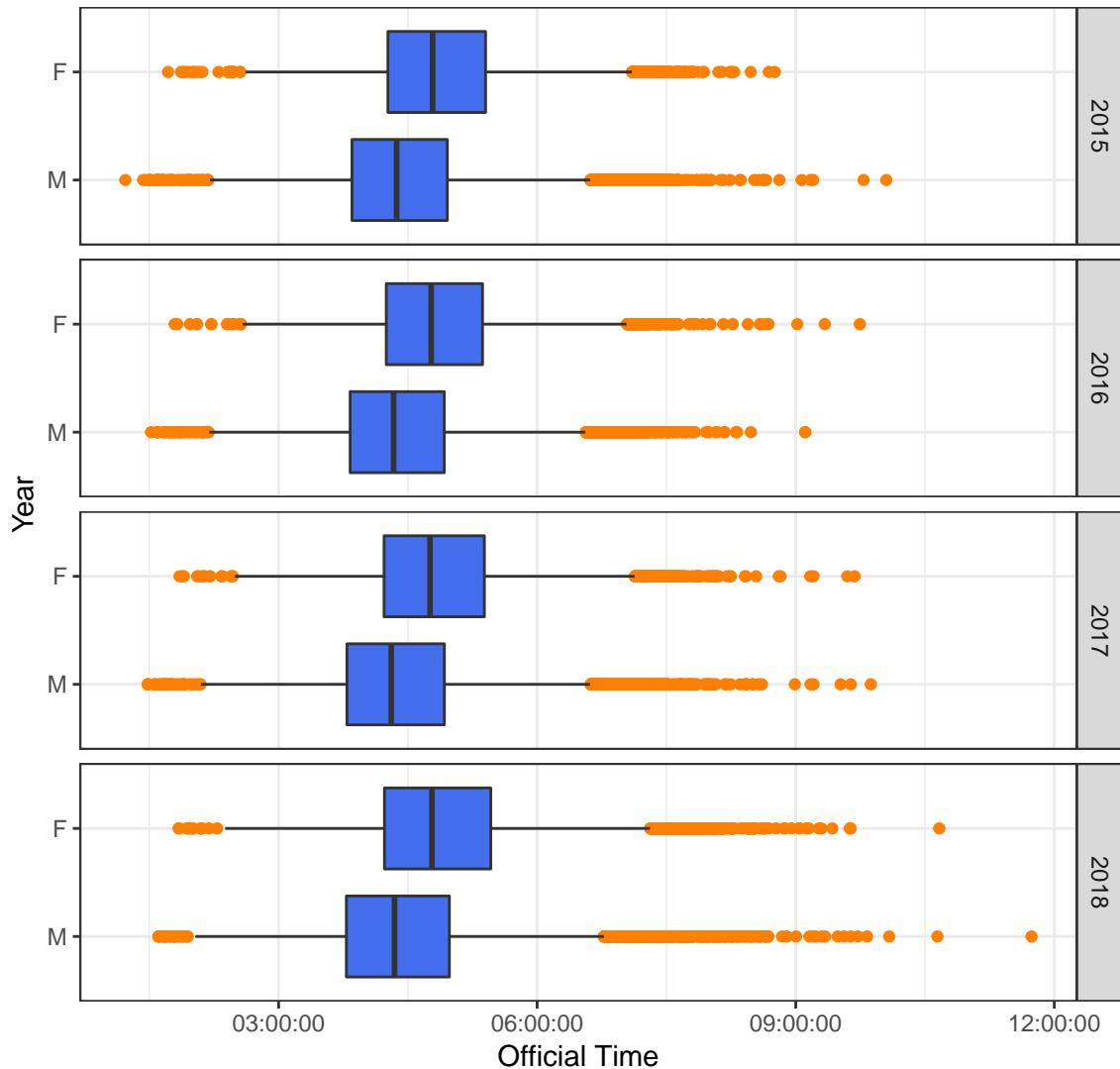
Is gender a significant indicator of performance? In all the following sections, we will only consider the runners, not other categories. Indeed, for instance, the wheelchair athletes would be considered as outliers since they race way faster than runners. Only considering runners will help us draw conclusions.

When analyzing a sport event, a natural question instantaneously arises: does gender affect performance, and to what extend?

Only take into account the runners, not the wheelchairs

Let us first draw a boxplot of the performance based on gender for the 4 year span we are analyzing.

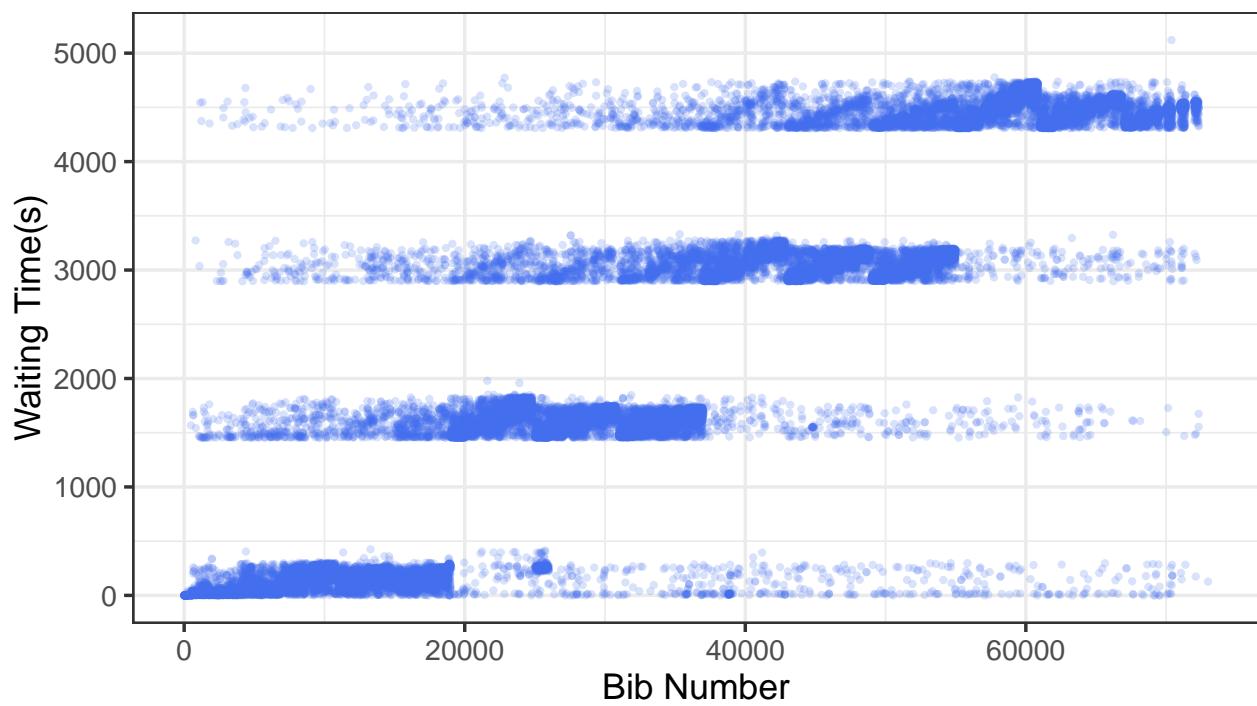
Each year, men perform better than women as a whole



We can clearly see that men consistently perform better than women in the marathon. For those four years, the women's first quartile is roughly the men's median. Therefore, given this plot, there is a clear correlation between gender and performance.

However, this simple plot does not take into account the runner's physical preparation. Let us now dive into the details of the race's organization.

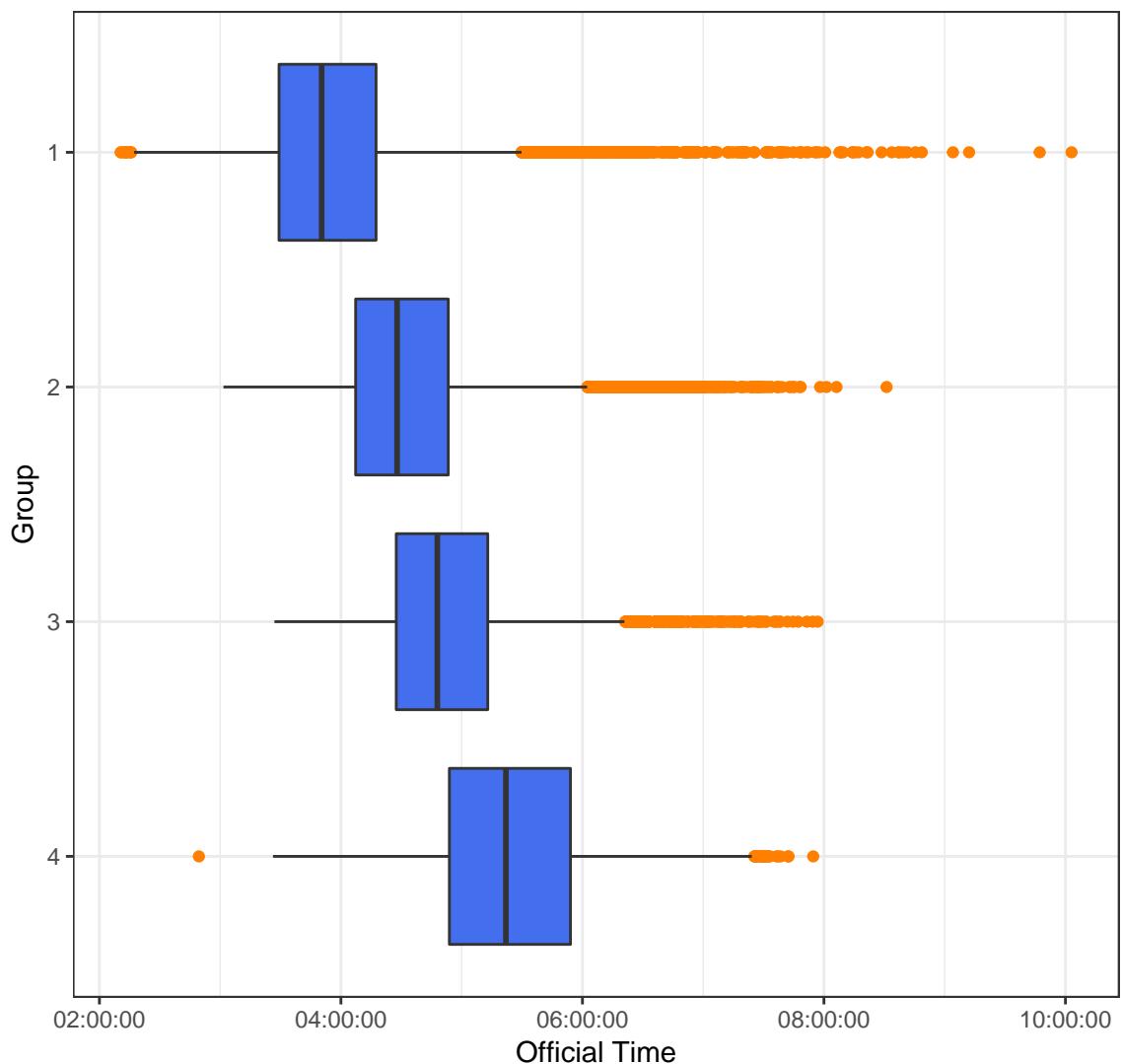
4 different groups clearly appear!



This first very simple plot gives us some information about the organization. First, there are clearly 4 different starts: * as expected, the first one starts at 0. It is the one that coincide with the gun start, * A second start occurs 25 minutes after the gun start, * A third start occurs 50 minutes after the gun start, * The fourth start occurs 1 hour 10 minutes after the gun start.

It is now clear that four different groups can be identified. But what are they based on?

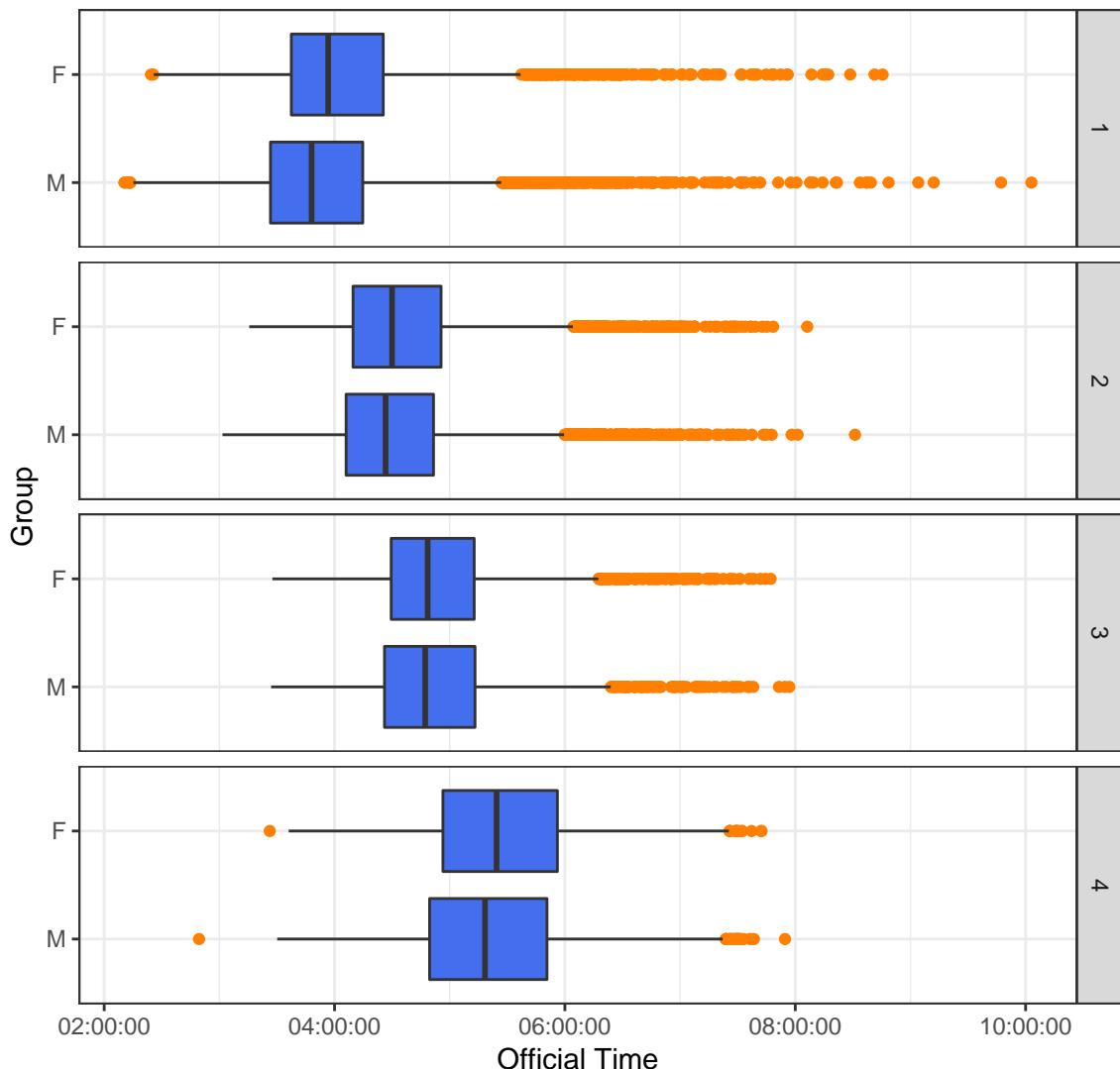
Groups are clearly based on performance



As we could have imagined, those groups are based on performance, the first one being the best one. In order to be allowed to run in a given group, participants have to prove that they already did a good enough time in an official race in the recent past. An obvious flaw of this method is that rookie runners cannot be placed in their real group since they have no record. We can see one of those in the fourth group: the outlier that run it in under three hours. Moreover, if a runner is out of shape but has a good record, (s)he still can be in the first group and still perform badly.

Now that we know how the groups are structured, we can look into them to see if there is a clear pattern acknowledging the fact that men perform better in marathon than women.

Different conclusions for each group



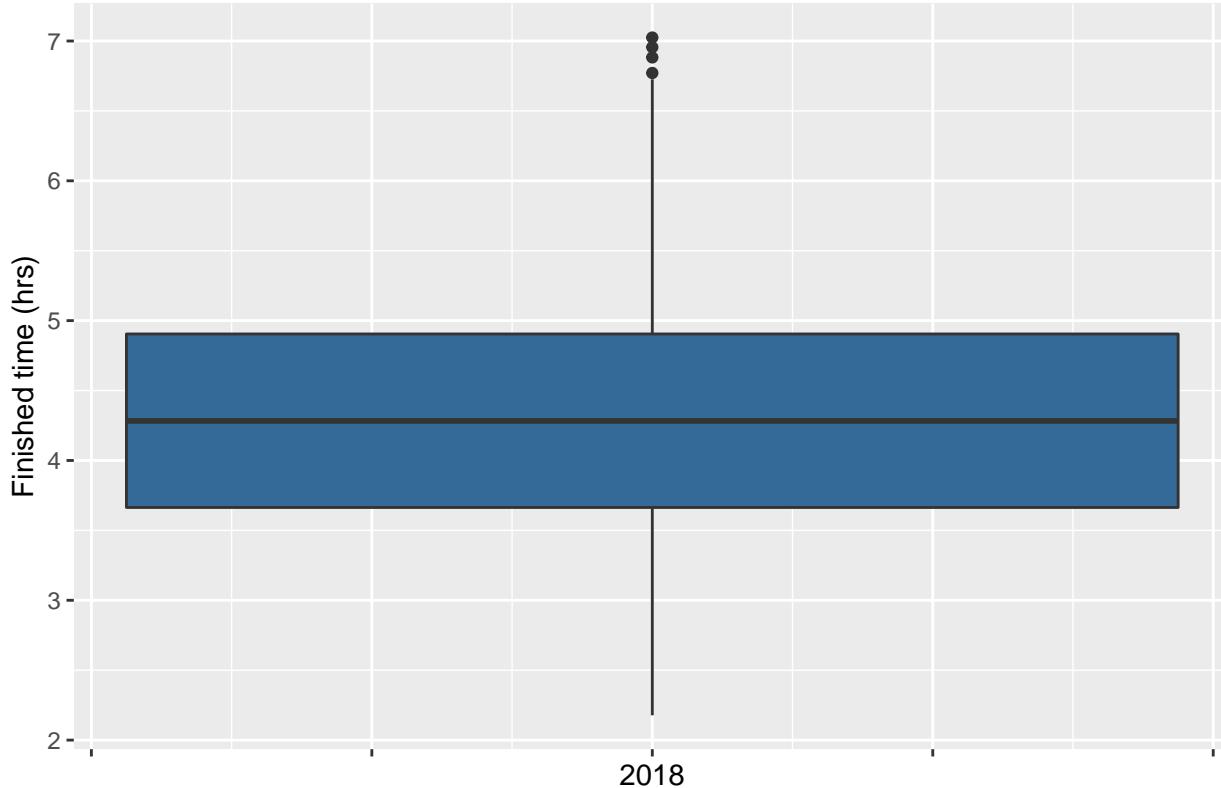
The results are quite striking and diverse: * **first group:** men significantly do better than women. * **second group:** men and women perform similarly. * **third group:** men and women perform similarly. * **fourth group:** men significantly do better than women.

How can we explain those two completely different behaviors of our data? It seems that for the runners in the intermediary groups, gender does not count when it comes to the race performance. Those runners have some prior experience and/or have trained regularly to achieve those results. It lets us think that **training can erase physiological differences**. The first group case is quite easy to interpret. **For athletes, training means pushing the body to its limits.** Therefore, it is not so surprising that this group exhibit a strong performance dependence on gender, since we find that dependence in almost all athletic disciplines.

4.3 Summarizing the whole race in a graph

We wanted to understand how we could summarize the race in a single graph. This was a challenge because we were unsure of what to graph to choose: should it be a histogram a box plot. Moreover, what variable should we used. The `official_time` but what about the performance between the different splits? Maybe some athlete had a strong start but finished poorly. Our first take to this problem was to generate a box plot of the running time.

Not informative



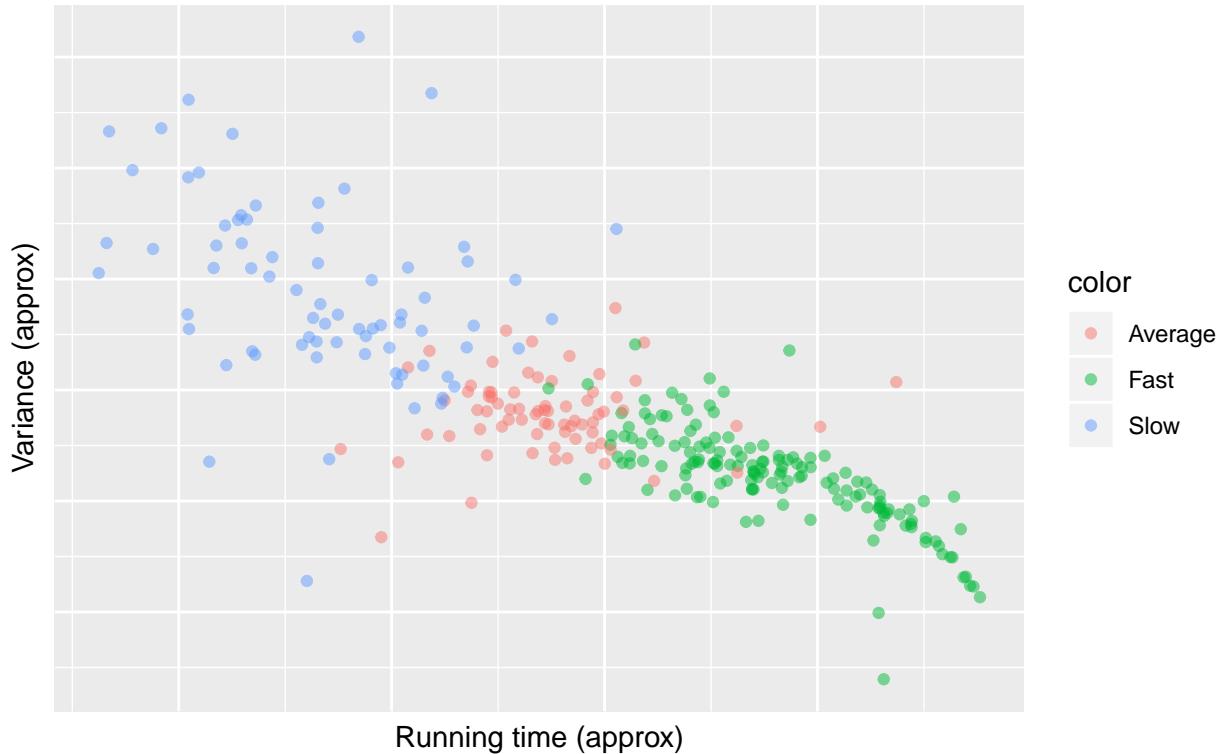
as we see above, the box plot is not as informative as we would like. It tells us that the finishing time was over 4 hours, where the mass concentrates (from slightly below 4 hours to 5 hours) and that the max and min. But we were not satisfied. Then, based on one of our ML homeworks, we saw that there was a technique that could go from distances between the observations to a two dimensional embedding. As an illustration, this technique could take the distances from all the cities within the US and recover the whole map.

where we see how New York and Boston are bunched together were as Seattle and Miami are thrown to two different corners. Thus, we thought that maybe this two dimensional embedding would be a useful summary of our data. The only missing ingredient was to define what the distance between different runners meant.

We found that the most useful summary resulted when we defined the distance between runners as the difference between each of its 5 k splits. The resulting embedding that we got for 2018 and only 10 % of the data is

Variance decreases as performance increases

Athlete Embedding by Running Time per 5k split (2018) – 10% sample



This is really cool! Before going into the details of the graph it is worth mentioning a couple of points. **The x and y axis are not the actual variables displayed in the labels but rather an approximation or interpretation** (it is like in PCA, where the two principal components are not a single variable but rather a combination of many that might render a certain interpretation). Thus, as seen from the graph **the x-axis displays the running time** where the time increases as you move to the left; the first green dot to the right was the fastest runner. It is worth pointing out that I colored the data points to show that this interpretation holds, but the embedding did what it thought best. Also, fast (green) means that your overall running time was below the median, average (red) means that you were above the median and below the 75% percentile and finally slow (blue) are the remaining points.

In terms of **the y-axis**, the interpretation is that this coordinate **captures the variance** in running time per 5 k split. Since points that had dissimilar performance on each split get separated, then the more dissimilar you are on each 5 k split to the athletes that had a similar running time, the further away you get pulled from them. As an illustration, take the green dot that is closest to the x-axis. That athlete had a great overall running time but the reason that it got thrown away from the elite “pack” is that it did not maintain a constant pace in all the splits as the rest of the top performers. What happened was that this athlete decreases severely its speed at the last 5 k split and that is why it got separated.

Given the introduction to this new type of graph, the main insights that we got are the following. Look how **the performance differences decelerate as we move to the right**. The graph resembles a logarithm curve that has been rotated on the x-axis. Based on this curvature, we see that a change for the green dots move you further above than in the red dots. **Thus, for the faster runners small perturbations in their running times sets them further apart than for slower runners.** This make a lot of intuitive sense since a difference in minutes is more detrimental and propagates a higher overall performance for the top athletes than for the rest. The second insight is that **the race was tight**. Look how close are the first points are. It looks as if they were on top of each other. Finally, we see something that we expected. **As performance decreases variance increases.** The green dots are close together because they maintain a constant pace and their finishing time was similar. In contrast, the slow runners kept changing pace on each

split and therefore the cloud of points is more spread around that.

In summary, we find the previous graph a powerful summary of the race. It tells you how tight the race was but also how much variance is there for the different running groups! Just for completeness we add the mathematical formulation. The problem that the embedding solves is

$$\min_x \sum_{i,j} (\|x_i - x_j\|_2 - D_{ij})^2$$

where D_{ij} stands for the difference in running times on each 5k split for runner i and runner j and the x s constitute the two dimensional embedding. Hence, we see how to minimize the previous objective function we have to make the difference in representation of each runner x_i and x_j match their actual distance D_{ij} .

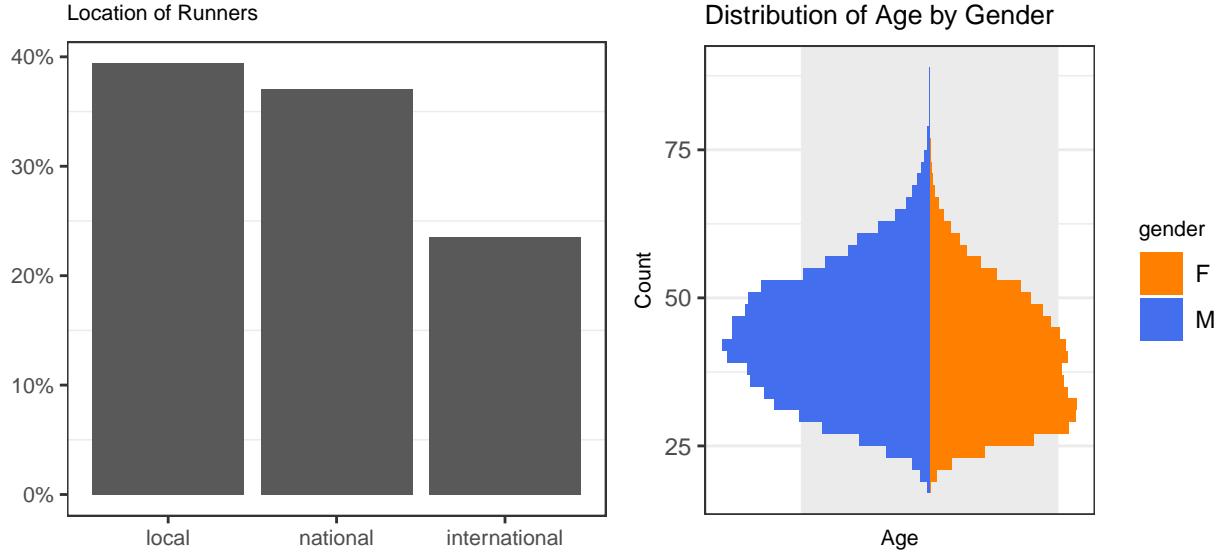
5. Executive Summary

Our analysis of the NYC Marathon data focused on answering the following questions (where we include a brief preamble).

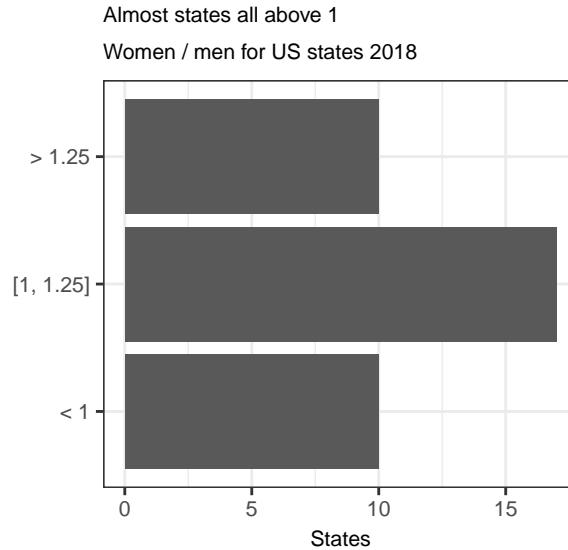
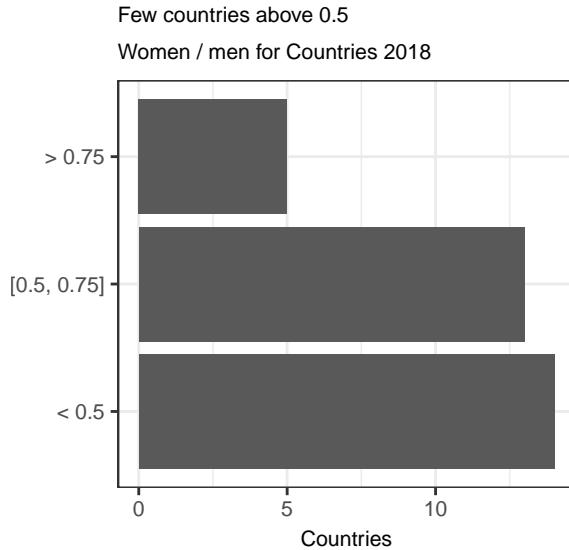
1. **Who are the people that run the NYC marathon and where do they come from?** The NYC marathon is one of the most important events in the city, thus we wanted to understand who ran this race. Is it only young people (30 and below)? It is mostly New Yorkers? Are there more men than women?
2. **How does the gender ratio compare across countries and states and what has been its trend overtime?** Due the overall inclusion of woman in various sports, we wanted to see how the NYC compare to this trend. Additionally, we wanted to see if woman inclusion was equal among the participating countries. Finally, we wanted to see its trend.
3. **Does gender matter in terms of performance?**. In some sports that have stringent physical requirements, men tend to perform better than women (for example, in Olympic weightlifting). However, we wanted to understand what was the gap in an endurance sport such as the NYC Marathon. Would woman always perform worse than men?
4. **How can we summarize the whole race in a graph?** Finally, placing demographics aside, we wanted to come up with a plot that summarized the whole race. How can we go about this? It is possible to *see* in a graph how tight the race was, how many odd events happened and how similar are the runners in all the race?

Next, we provide an answer to each of the previous questions.

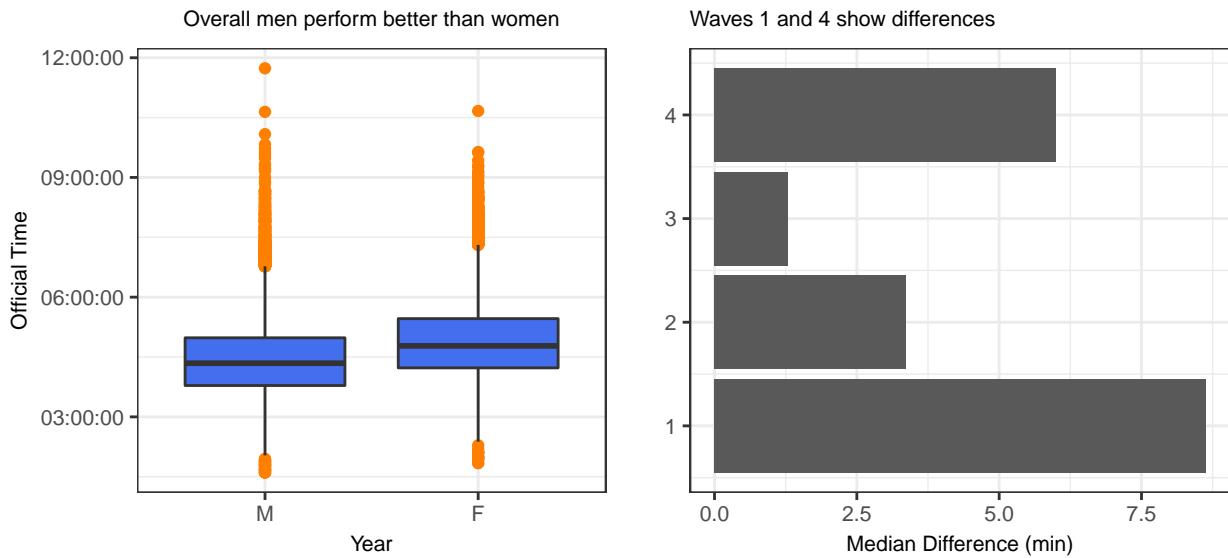
1. In terms of location, **we find that athletes mostly come from the US (over 70%) and the rest are international**. In terms of age and gender, **we found that the majority of the participants are young adults around 40 years old and that, on average, female athletes tend to be younger**. Also, we say that as age increases the number of female participants decreases.



2. We found that the gender ratio is generally low and that only 3 countries are above 0.8 (women / men). Nonetheless, in the majority of the US states have an above 1 ratio. Please see section 4.1 for the full decomposition of each graph.



3. We found that the difference in performance between women and men is only observed in groups that rely on more capacity (for full details see section 4.2).



4. Finally, we saw that we could summarize the whole race by using an embedding. This embedding allowed us to visualize (1) how close the race was by looking at the distance between points, (2) what is the variance of the data set by looking at the spread on the y-axis, (3) if odd events happened by noticing the dots that were apart from their closest group. See section 4.3 for all the details

6. Interactive component

Interactive component can be found in shinnyapps server ([link](#))

For the interactive component, we decided to generate a dashboard that could let us summarize the information for each location (states or countries), and compare some results with the overall, across the four years of marathon data we collected. We believe that summarizing data in a graphical way

Creating this App will allow users to choose between countries and states across time, and compare different metrics:

- The gender-age distribution
- The proportion of woman and men and compare it with the rest of the world/states
- The official time distribution of a specific location compared with the overall
- The change in terms of performance among athletes, as race advances the distribution widens and for some countries this becomes more clear than for others.

Having summarize information by country and state can be very useful for people interested in marathons, sponsors and even governments to try to promote the participation, specially in countries with not many participants but performing above average (as the case of Kenya). As well, it can encourage the a more equal participation between woman and men by comparing their country/state with the rest of the world/country. We expect that the interactive component of this project can help us understand more about the people participating and the differences in terms of representation and performance compared to the total.

To use this tools one should choose a specific year and location from the sidebar. The location can be to choose between a state from the US or a country. Since there were too many options, we use some shiny Widgets, that do a live-search among the options when typing the first letters.

Some of the key findings in this interactive was looking at the different results, specially for countries like Kenya with a distribution of official times outperforming the general or Taiwan on the opposite side.

7. Conclusion

Many of the problems that we faced were found in the web scrapping part of the process. As mentioned in previous sections, we had to overcome several challenges in learning how to properly scrape the NY Road Runners website including parsing the webpage and setting up the computer engine. Next came cleaning the data and dealing with messy locations, time format, and making decisions about dealing with missing data. For example, many of the rows needed to be dropped because we could not identify the country in which the city or county resided for that particular runner. This exclusion could have impacted the outcome of our analysis as many city names in the US and Canada are identical to ones found across Europe. Another issue that we had to make a decision about was the handling of data concerning the wheelchair racers and hand-cyclists. There were very few of these participants in comparison to the runners, much to few to dedicate worthwhile insights, and thus those two groups were largely ignored throughout the analysis of performance.

One of the lessons that we learned was the importance of analyzing our data across different variables as opposed to only examining each variable individually. For example, by examining proportion of men and women at the country level, we discovered more information about country-level and state-level trends that we would otherwise explored. More generally, we learned that it is important to be curious and creative with the research directions that we chose in order to extract the most interesting pieces of analysis.

Had there been more time to dig deeper in this analysis, it would have been interesting to add more years in our analysis to analyze trends and shifts in the overall number of people participating, the ratio of women to men, the international representation, and perhaps most interestingly the trends in performance. Has the NYC Marathon become more or less competitive over the years? Another interesting question to look at would be any changes in the geography of the race and how that impacted performance over time? It is a fact that NYC contains a lot of hills that pose challenges to marathon runners, and it is possible that changes in the format of the race could have a significant impact on performance. Finally, a more extensive analysis of the teams participating would give us more insights into how professionals train for a race, and what has been the trend in the impact, if at all, of being a part team on performance. In our analysis we found that teams did not have an effect on the outcome of individual runners, but this could be part of a longer, more dynamic trend. For example, we might discover that many years ago only the professional athletes could train for a marathon successfully, while the amateur solo runners were not able to be competitive.

In summary, we enjoyed completing this analysis and learning more about a famous NYC tradition, and hope to be a part of a data set on marathon finishers one day soon!