STATS 15 FINAL - Track and Field

Casper Hsu, Jason Liu, Anthony Chen, Aryan Mosavian Pour, Amy Zeng

2022-11-30

Motivating Questions

Track and field has a lengthy history of controversy when it comes to technological advancements. Runners and organizations regularly debate whether recent technological advancements ruin the integrity of the sport as they provide too much of an aid to racers. We decided to conduct an exploratory analysis of 21st century elite running results and the relations between observed trends and technological advancements in the sneaker market. In specific, we ask: ***"can sudden downward trends in track and field running marks be attributed to advancements in track and field technology over the past 22 years?"***

The Dataset

For this project, we will be working with a dataset of the top 100 track results for each year by event. It's publicly available on the website of the World Athletics, the official governing body of the sport of Track and Field. It contains comprehensive records from each year since 2001 (with a few exceptions).

Web Scraping

However, the database provided has each event and year combination within a different web address (example: https://www.worldathletics.org/records/toplists/sprints/100-metres/outdoor/men/senior/2001). This means that in order to create a single dataset, we must scrape the tables of each html page and compile them into a single csv.

For this, we decided to use the Python packages BeautifulSoup and urllib3 since they are extremely well documented for similar applications. We wrote a simple script to concatenate url's to find the location of each table and pull the data from each field, appending it to a new csv file named "Top100TrackResults.csv".

This almost completely maintained the original structure of the dataset. The only things that we had to change were: for entries of events where wind isn't recorded, a null field was added in place of it to maintain a constant amount of columns and since the website's formatting method added an inconsistent amount of spaces to each field, it was best to just remove all spaces in the entries.

The webpage for the Men's Marathon results in 2018 was inexplicably missing, but in order to be consistent with our data source, we chose to leave it omitted. While not missing, results for 2020 appear consistently not representative of skill levels at the time due to there being almost no major races during the pandemic, however, we decided to not remove them until we can conclusively justify it in a later step.

We have extensively cross-referenced our data set with some other reputable sources to verify that our information is accurate and complete.

Background Information

Modern Track (formerly known as athletics) is based on the footraces of the Ancient Greeks as a measure of running skill. Due to it's ease of access, it's popular in almost all parts of the world.

For each race:

- Competitors line up on the start line
- An Official fires the starting gun, starting a digital timer
- Upon hearing the sound, Competitors race to run the specified distance in as little time as possible
- After completing the laps/distance, the first Competitor who's chest crosses the plane of the finish line is declared the winner and awarded "1st". The following runners are awarded "2nd", "3rd", "4th", etc.. as they cross.
- The time measured between the gun going off and each runner finishing is recorded. For 100m and 200m races, the wind is also measured and factored in to create adjusted times (the time they would have achieved without the wind)

There are many events in Track, but for this report, we'll only focus on the more relevant ones:

- Marathon: Competitors run 26.2 miles on the road with varying elevation and course shapes
- Half Marathon: Competitors run 13.1 miles on the road with varying elevation and course shapes
- 10,000m: Competitors run 25 laps (6.25 miles) on a regulation track
- 5,000m: Competitors run 12.5 laps (3.125 miles) on a regulation track
- 1500m: Competitors run 3.75 laps (0.932 mile) on a regulation track
- 800m: Competitors run 2 laps (0.5 mile) on a regulation track
- 200m: Competitors run 1/2 lap (0.125 mile) on a regulation track
- 100m: Competitors run 1/4 lap (0.0625 mile) on a regulation track

While all these events involve running, they consist of different types of running. Longer events (Marathon, Half Marathon) rely mostly on an athlete's ability to efficiently store and utilize energy reserves, while short events, known as sprints (100m, 200m), require a high peak power output. However, it's more of a spectrum since events between utilize elements from both. Distance (10,000m, 5,000m) focuses on endurance and stamina, while Mid Distance (1500m, 800m) balances both strength and power.

Due to its heavy reliance on pure athleticism (lung capacity, leg strength) rather than learned skills in other sports (throwing, dribbling), most improvements in track come rather slowly. Long Distance runners train in extended cycles, carefully timing "workouts" (consisting of high effort intervals) and "volume" (training by running higher mileage) to make minuscule gains for their important "peak races" (often the Olympics or major title competitions). Better understanding of how to optimize these cycles, along with increased of the function of the human body has led to gradual improvements in time. In theory, these should level out as they approach the asymptotic "theoretical limit", the fastest a human being could ever cover the specified distance. Some speculate that 100m runners are already very close to it, given that there is simply less to optimize within a 9 second race.

As a result, many look to technology as a way to quickly improve results and push the limits of what humans are capable of. At the elite level, athletes are constantly looking to get an edge over the competition or prevent themselves from falling behind. One of the most relevant and controversial running technologies are so-called "supershoes".

Runners have been trying to get an advantage through footwear since at least the early 70s, when Bill Bowerman, Head Coach at University of Oregon and co-founder of Nike, created custom shoes for his athletes using a waffle iron.

Being thinner, lighter, and having superior traction, they logically would give any athlete using them a slight advantage over the competition. Once brought to market, constant competition with other manufacturers led to the creation of thinner and thinner shoes, slowly reducing the amount of rubber to be lighter each year. For the track, they developed needle-like pins on the bottom to provide optimal traction on the rubberized surface. For the road, they reduced protrusions to optimize surface area contact on the asphalt under varying conditions.

However, in 2016, Nike shook this up by announcing their Breaking2 project, where they dedicated a whole team of scientists to try to break the (often deemed impossible) 2 hour marathon barrier. They selected 3 of the world's best marathoners, and while they made some minor improvements in fueling and training methods, most of the improvement could be traced back to the unorthodox shoes they had developed throughout the project.

Progression of Marathon Shoes

Marathon shoes prior to this were engineered to be as thin as possible while still providing just enough support to finish the distance.

2016- Zoom Streak 6

At the time, the best racing flat for Marathons was Nike's Zoom Streak 6. It was particularly thin and light, with good ground-feel. It had a stack height of only 26mm and weighed 181 grams.

2017- ZoomX Vaporfly 4%

After being used exclusively by the top Nike athletes throughout 2016, the now infamous Vaporfly 4% was released to the public in 2017. However, it wasn't initially received well due to its seemingly backwards design. While all other shoes had a minimalist design approach, this did the exact opposite. It had a 39mm stack height and weighed 198 grams. The shoes were high off the ground and unstable, with a profile that tapered straight down. Many elite marathoners refused to use them due to how unwieldy and awkward they were.

However, they were still revolutionary, in their own way. They used a new lower density foam known as Pebax which proved to be more energy efficient than any other foam by far. The most debated element, however, is the use of a carbon fiber plate within the midsole. It supports the foot throughout each step and returns energy when pushing off.

An independent study by the University of Colorado Boulder (https://www.colorado.edu/today/2018/11/20/what-makes-worlds-fastest-shoe-so-fast-new-study-provides-insight) verified Nike's claim made in the name, that the new shoes made runners 4% more efficient. Seeing this, other companies started to develop their own supershoes, adapting the key aspects of Nike's design to fit the niche of their brand, therefore starting the supershoe revolution. Due to their massive head start and their unique development methods, Nike has remained the leader, so we will mark technological improvements by Nike's releases. This is because elite runners have begun to quickly adopt whatever is the newest technology, only switching to the other brand's offerings if it suites them better, but the general standard of technology is always dictated by Nike's latest and greatest shoe. It's even become a common practice for athletes sponsored by other companies to paint Nikes to pass them off as the other brand and compete in them (https://www.businessinsider.com/nike-vaporfly-shoes-runners-with-other-sponsors-wear-secretly-2020-8).

However, in 2017, many elite marathoner's chose to remain with their standard racing flats, not completely trusting the newer supershoes and having many concerns over the stability and reliability of them. Below is a chart of the shoes used by podium finishers in the major marathons of 2018.


2018 WORLD MARATHON MAJORS
TOP 3 SHOES
RUNNING SHOES GURU
1
2
3
TOKYO MARATHON
MEN
DICKSON CHUMBA
2:05:03
STREAK 6
YUTA SHITARA
2:06:11
VAPORFLY 4%
AMOS KIPRUTO
2:06:33
ADIZERO SUB2
WOMEN
BIRHANE DIBABA
2:19:51
VAPORFLY 4%
RUTI AGA
2:21:19
VAPORFLY 4%
AMY CRAGG
2:21:42
VAPORFLY 4%
BOSTON MARATHON
MEN
YUKI KAWAUCHI
2:15:58
CUSTOM SORTIE MAGIC
GEOFFREY KIRUI
2:18:23
CUSTOM VAPORFLY 4%
SHADRACK BIWOTT
2:18:35
PROTOTYPE HYPERION
WOMEN
DESIREE LINDEN
2:39:54
PROTOTYPE HYPERION
SARAH SELLERS
2:44:04
1400 V5
KRISTA DUCHENE
2:44:20
FASTWITCH 8
LONDON MARATHON
MEN
ELIUD KIPCHOGE
2:04:17
VAPORFLY ELITE FLYPRINT
KITATA SHURA
2:04:49
VAPORFLY 4%
MO FARAH
2:06:21
VAPORFLY ELITE
WOMEN
VIVIAN CHERUIYOT
2:18:31
VAPORFLY 4%
BRIGID KOSGEI
2:20:13
STREAK 5
TADELECH BEKELE
2:21:40
VAPORFLY 4%
BMW BERLIN MARATHON
MEN
ELIUD KIPCHOGE
2:01:39 WR
VAPORFLY ELITE VER. 2
AMOS KIPRUTO
2:06:23
ADIZERO ADIOS
WILSON KIPSANG
2:06:48
ADIZERO SUB2
WOMEN
GLADYS CHERONO
2:18:11
ADIZERO ADIOS
RUTI AGA
2:18:34
VAPORFLY 4% FK
TIRUNESH DIBABA
2:18:55
VAPORFLY 4% FK
Bank of America Chicago Marathon
MEN
MO FARAH
2:05:11
VAPORFLY ELITE VER. 2
MOSINET GEREMEW
2:05:24
VAPORFLY 4% FK
SUGURU OSAKO
2:05:50
VAPORFLY 4% FK
WOMEN
BRIGID KOSGEI
2:18:37
VAPORFLY 4% FK
ROZA DEREJE
2:20:49
VAPORFLY 4% FK
SHURE DEMISE
2:21:46
ADIZERO ADIOS
TCS NEW YORK CITY MARATHON
MEN
LELISA DESISA
2:05:59
VAPORFLY 4% FK
SHURA KITATA
2:06:01
VAPORFLY 4% FK
GEOFFREY KAMWOROR
2:06:26
VAPORFLY ELITE VER. 2
WOMEN
MARY KEITANY
2:22:48
ADIZERO TAKUMI SEN
VIVIAN CHERUIYOT
2:26:02
VAPORFLY 4% FK
SHALANE FLANAGAN
2:26:22
OG VF4% X VF-ELITE
RUNNING SHOES GURU

While many of the top runners had adopted the shoes, some podium finishers and a large amount of the remaining field still opted for traditional designs like the Adidas ADIZERO offerings. This graphic demonstrates the significant edge Nike athletes were getting over their competition a midst a large variety of standard picks. Note that this yearly chart wasn't produced until 2018 to track the growing prevalence of supershoes since not a relevant amount of runners used them in 2017.

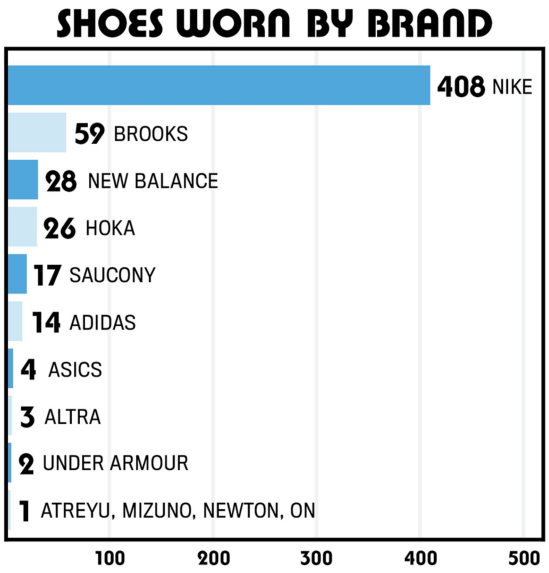
2019- ZoomX Vaporfly Next%

Featuring an upper with better lock down and a wider forefoot that flared out, the (now nicknamed) Next% fixed many of the stability issues of the previous generation. Even if it didn't give any mechanical advantage, the increased comfort and support of something essentially the same weight as a traditional racing flat gave athletes no reason not to use them.


RUNNING SHOES GURU
2019 WORLD MARATHON MAJORS
RUNNING SHOES OF THE FIRST 3 MEN AND WOMEN
TOKYO MARATHON
MEN
BIRHANU LEGESE 1ST 2:04:48 VAPORFLY 4% FLYKNIT
BEDAN KAROKI 2ND 2:06:48 VAPORFLY 4% FLYKNIT
DICKSON CHUMBA 3RD 2:08:44 VAPORFLY 4% FLYKNIT
WOMEN
RUTI AGA 1ST 2:20:40 VAPORFLY 4% FLYKNIT
HELEN TOLA 2ND 2:21:01 VAPORFLY 4% FLYKNIT
SHURE DEMISE 3RD 2:21:05 ADIZERO ADIOS 4
BOSTON MARATHON
MEN
LAWRENCE CHERONO 1ST 2:07:57 VAPORFLY 4% FLYKNIT
LELISA DESISA 2ND 2:07:59 VAPORFLY 4% FLYKNIT
KENNETH KIPKEMOI 3RD 2:08:07 VAPORFLY 4% FLYKNIT
WOMEN
WORKNESH DEGEFA 1ST 2:23:31 ADIZERO ADIOS 3
EDNA KIPLAGAT 2ND 2:24:13 VAPORFLY 4% FLYKNIT
JORDAN HASAY 3RD 2:25:20 ZOOM VAPORFLY 4%
LONDON MARATHON
MEN
ELIUD KIPCHOGE 1ST 2:02:37 ZOOMX VAPORFLY NEXT%
MOSINET GEREMEW 2ND 2:02:55 ZOOMX VAPORFLY NEXT%
MULE WASIHUN 3RD 2:03:16 ZOOMX VAPORFLY NEXT%
WOMEN
BRIGID KOSGEI 1ST 2:18:20 ZOOM VAPORFLY 4%
VIVIAN CHERUIYOT 2ND 2:20:14 ZOOMX VAPORFLY NEXT%
ROZA DEREJE 3RD 2:20:51 ZOOMX VAPORFLY NEXT%
BMW BERLIN MARATHON
MEN
KENENISA BEKELE 1ST 2:01:41 ZOOMX VAPORFLY NEXT%
BIRHANU LEGESE 2ND 2:02:48 ZOOMX VAPORFLY NEXT%
SISAY LEMMA 3RD 2:03:36 ZOOMX VAPORFLY NEXT%
WOMEN
ASHETE BEKERE 1ST 2:20:14 ZOOM VAPORFLY 4%
MARE DIBABA 2ND 2:20:21 ZOOMX VAPORFLY NEXT%
SALLY CHEPYEGO 3RD 2:21:06 ZOOMX VAPORFLY NEXT%
BANK OF AMERICA CHICAGO MARATHON
MEN
LAWRENCE CHERONO 1ST 2:05:45 ZOOMX VAPORFLY NEXT%
DEJENE DEBELA 2ND 2:05:46 ZOOMX VAPORFLY NEXT%
ASEFA MENGSTU 3RD 2:05:48 ZOOMX VAPORFLY NEXT%
WOMEN
BRIGID KOSGEI 1ST 2:14:04 ZOOM VAPORFLY 4%
ABABEL YESHANEH 2ND 2:20:51 ZOOMX VAPORFLY NEXT%
GELETE BURKA 3RD 2:20:55 ZOOMX VAPORFLY NEXT%
TCS NEW YORK CITY MARATHON
MEN
GEOFFREY KAMWOROR 1ST 2:08:13 ZOOMX VAPORFLY NEXT%
ALBERT KORIR 2ND 2:08:36 ADIZERO PRO
GIRMA BEKELE GEBRE 3RD 2:08:38 ZOOMX VAPORFLY NEXT%
WOMEN
JOYCILINE JEPKOSGEI 1ST 2:22:38 TAKUMI SEN 5
MARY KEITANY 2ND 2:23:32 ADIZERO PRO
RUTI AGA 3RD 2:25:51 ZOOMX VAPORFLY NEXT%
RUNNING SHOES GURU
WWW.RUNNINGSHOESGURU.COM

While other picks were fairly common prior to it, after the Next% release during the 2019 London Marathon, the whole field wore the shoe. The only exceptions to this were a few Adidas sponsored athletes using Adidas' version, the ADIZERO Pro, and Joyciline Jepkosgei, a 10,000m prodigy, running her first Marathon ever.

By the end of 2019, the shoe choices of the top 500 American Marathoners were as follows:



The iconic neon green (and later neon pink) colored shoes became so prominent that upon claims of being unfair, World Athletics was pressured to update their shoe regulations. To prevent invalidating over a year's worth of records, they decided to freeze the maximum specs where they were, allowing a non-overlapping rigid body and a maximum stack height of 40mm (https://www.worldathletics.org/download/download?filename=b723c6b6-7d1f-40ad-8b27-1d3f956c6c99.pdf&urlslug=C2.1A%20%E2%80%93%20Athletics%20Shoe%20Regulations%20(effective%20from%2001%20January%202022)).

2020- Air Zoom Alphafly Next%

After continually missing the 2 hour barrier, top Marathoner Eliud Kipchoge, one of the main testers of the Vaporfly, developed a new version using Zoom Air, Nike's patented technology of pressurized air units filled

with tensile fibers and put it into the Alphafly Next%.

While this feat was extremely impressive, the shoe failed to see widespread use in a competitive environment. The shoe was bulky and no one wanted to use it since it was so high off the ground.

As a result, the shoe never took off. Even upon its launch at the 2021 US Olympic Trials, when many of the above flaws weren't well known, it was picked less than 30% of the time. After that race, it dwindled off as a one-off pick, with mostly just Kipchoge and a few others using it.

Kipchoge even had a bad race with it. Citing a cramp affecting his form and being further magnified by the instability of the shoes, he finished 8th in the 2020 London Marathon with a time of 2:06:49 (many minutes worse than normal) and losing for the first time since 2013. He ended up switching back to the Vaporfly for the next 2 years before picking them up again in 2022.

Progression of Mid Distance Spikes

The design and construction of track spikes is particularly important for middle distance events due to the extreme conditions runners are exposed to. Runners average speeds well above 15mph and can temporarily reach above 20mph while accelerating to contend for spots. This, along with the fact that the pace must be maintained for a matter of minutes, means that spikes must be constructed to provide sufficient support to properly sprint while also being light and efficient. The general solution throughout the 2000s was to create a thin hybrid spike where the front is a semi-rigid sprint plate, but the back is flexible and cushioned.

2008- Zoom Victory

The Nike Zoom Victory was the classic bare-bones middle distance spike. Material was placed only where absolutely necessary, with a thin external plastic plate only on the forefoot. It's feather-like weight made it a good choice over the offerings at the time.

2012- Zoom Victory Elite

The Victory Elite, however, took a different approach. Instead of eliminating elements to be lighter, it added a second plate, this time made of carbon fiber that ran about 3/4 the length of the shoe. This was effective because the increased rigidity in the midfoot supported the foot and kept the runner on their forefoot, even when tired. Forefoot running is widely accepted as more energy efficient and allowing for a snappier push off, facilitated by the increased stiffness of the shoe. This was the first example of the use of carbon fiber significantly improving the performance of a shoe.

2016- Zoom Victory Elite 2

The next iteration used a very unique method to create the spike plate. Nike started with the lightweight structures found on coral to create an organic honeycomb shape. Then, they used force data from top sprinters along with various physics simulations and an AI algorithm to create many iterations, each of which were rapidly prototyped using 3d printing and tested (https://www.engadget.com/2016-08-03-nike-zoom-superfly-elite-3d-printing-olympics.html). This created a spike plate with unparalleled sprint performance, as each vertex would bite into the ground at the optimal depth and the mechanism would flex and delivery energy as efficiently as possible through the pliant mechanics achieved by the varying dimensions and wall thickness of each cell.

It also included the carbon plate, but it was a little more flexible and the main characteristics of it were hard to distinguish from the front-heavy 3d spike plate. While they were certainly better for sprinting, the increased resistance and strong sprint characteristics made them less than ideal for later stages of the race, leading many elite athletes to favor the original Victory Elite.

2021- Air Zoom Victory

This spike brought applied many of the technical elements developed for Marathon shoes to the track. Instead of using a rigid 3/4 length plate like before, it adopted the internal full length pliant plate from the Vaporfly with more of an aggressive toe-off characteristic. A large unit of Zoom Air is sandwiched between that and a low-resistance external spike plate. This puts the design at the maximum legal specifications for track spikes: having a single internal and external plate while remaining below 20mm. The Air unit mounted directly onto the sweet-spot of the carbon plate creates a fast rebound, generating a torsion force on the external plate, forcing the runner both up and forwards. This, along with the light weight from the minimal use of Pebax foam and a thin steamed knit upper, creates a tremendously efficient spike that is tremendously responsive for sprinting.

Since its release, its become widely accepted as the best choice for middle distance, particularly 800m.

(Late) 2021- ZoomX Dragonfly

Introduced late in the 2021 season and not seeing significant use until 2022, the Dragonfly is another track spike utilizing marathon technology. However, it has only a single nylon plate, which acts as both the internal and external. The lack of Zoom Air in it and use of a more flexible plastic leads it to be far less responsive. While this makes it undesirable for the speeds achieved in an 800m, since it stays more grounded and has a smoother transition, some believe it is advantageous for the 1500m, leading it to share the top pick with the Air Zoom Victory in 2022.

Progression of Sprint Spikes

Since sprinting has almost no energy component, sprint spikes aim to improve traction and support (particularly through rigidity) at the cost of weight.

2012- Zoom Superfly R4

Following traditional design principles, the bottom of the spike consisted of a large rigid plastic piece, supporting as much of the foot as possible. Not much was known about power delivery and biomechanics at the time, so most design choices followed the assumption that stiffer is better.

2016- Zoom Superfly Elite

Using the same 3d traction technology as the Victory Elite 2, the data-driven design of the Superfly Elite proved to have superior traction and power delivery. The angles support the footstrike and increased stiffness where needed optimizes the direction and timing of the release. Spikes of this design dominated the 100m and 200m distances.

2021- Air Zoom Maxfly

Based on the Air Zoom Victory, the Maxfly instead has a mostly rigid carbon plate and a smaller Zoom Air unit (shifted further towards the front). The external plate runs the full length of the shoe and it uses standarad pylon instead of Pebax, making it compress and flex much less. This achieves a more stable feel while still keeping most of the responsiveness of the Victory. However, to stay under the stack height limitations, they couldn't use 3d traction, opting for a similar external plate to the Victory but with additional spiked ridges. Throughout 2021 and 2022, this has been the dominant pick for sprinters due to how responsive it feels underfoot.

Exploration of Variables

Explanatory Variables:

- Gender (categorical): female (W) or male (M) runner. Due to the uniformity collection of our data we selected, the number of men and women should be the same. This is since we take the top 100 for each year, for each event for each gender.
- Distance (categorical): we analyzed 8 different run distances to see the impact that length of race would have on the amount of change that technology had on the time it took to run them. Due to the uniformity collection of our data we selected, the number of distances should be the same. This is since we take the top 100 for each year, for each event, for each gender.
  - Road (Marathon/Half Marathon)
  - Distance (10000m/5000m/3000m)
  - Middle Distance (1500m/800m)
  - Sprints (400m/200m/100m)
- Rank (numerical): the position of the runner from 1 to 100; there may be ties. Since the dataset is the top 100 per year, the distribution of ranks should be the same, with 21 values of each rank. (positive integer)
- Wind (numerical): wind speed measured using an anemometer; can be both positive and negative to denote headwind and tailwind in meters/second. Wind only applies to the 100, and 200m races. Since it is factored into the Mark variable, it should have low to no impact. (positive real number)
- Competitor (categorical): name of the runner competing. It is formatted in FirstLAST
- DOB (categorical): date of birth of the runner in day, month, and year order. Each competitor has a distinct date of birth.
- Nat (categorical): nationality of the runner; expressed using 3 letter abbreviations. Doing analysis on nationalities raises many ethical concerns so we have decided not to include it.
- Pos (numerical): position of runner. The position of the runner is distinct for each race and each race is independent so there will be multiple number 1 position. Only so many people can get a certain place due to the limited number or races. (positive integer)
- Venue (categorical): country and stadium that the race took place in. The venues introduce many confounding variables dealing with weather and altitude so it will not be used in our analysis. Many races will be held at seasonal competitions' venues such as the olympic stadiums or world championships.
- Date (categorical) (correlates to technology): date that the race took place on in day, month, and year order. Each competitor can only show up in the top 100 once per year, so it will be an even spread of dates.
- We took 2020 out because it is an outlier (explained later)
- ResultsScore (numerical): arbitrary measure that the governing body uses to score the season performance (factors like difficulty of race); in the 1100 to 1200 range. It is made up of many factors that are already included in the dataset using an algorithm created by World Athletics. (positive integer)

Response Variable

- Mark (numerical): seconds it takes for the runner to complete the designated distance. Each race has a distinct mark range.

Data Cleaning

Installing packages and dataset csv. The packages that are used are the tidyverse, and the chron package to help clean out times into a consistent format

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(chron)
```

```
##
## Attaching package: 'chron'
##
## The following objects are masked from 'package:lubridate':
##
##     days, hours, minutes, seconds, years
```

```
library(rmarkdown)
library(dplyr)
results <- read.csv("Top100TrackResults.csv")
```

Arguably the most important variable to us is the Mark variable, which represents the time recorded between the runner leaving their mark and crossing the finish line. Values in the master data set are stored in either the format SS.MS (Seconds, Milliseconds), MM:SS (Minutes, Seconds), or HH:MM:SS (Hours, Minutes, Seconds). Events that have a value for the hour placeholder are all Marathon times. To make data wrangling and comparing times between marathons and other distances easier, we created the Time variable.

```
h_first <- filter(results, Distance == "Marathon") %>%
  mutate(Time = chron(times. = Mark))
```

Since half-marathons are commonly finished reasonably close to the hour mark, there are a considerable amount of half-marathons recorded in our dataset that finished under the 1 hour mark, and an even larger amount that finished over the hour mark. This raises a problem because times under 1 hour are recorded as MM:SS, whereas times over 1 hour are recorded as HH:MM:SS. To solve this issue, we simply keep the HH:MM:SS times as is, and paste a 00: before the MM:SS times to make them the same format.

```
h_first_half <- filter(results, Distance == "Half Marathon", nchar(Mark) == 7) %>%
  mutate(Time = chron(times. = Mark))
```

```
m_first_half <- filter(results, Distance == "Half Marathon", nchar(Mark) == 5) %>%
  mutate(Time = chron(times. = paste("00:", Mark, sep = "")))
```

Since 10k, 5k, 1500m, and 800m races are all also recorded in the MM:SS format, we will use a similar procedure to reformat them into the more comparable HH:MM:SS

```
m_first <- filter(results, Distance %in% c("10000m", "5000m", "1500m", "800m")) %>%
  mutate(Time = chron(times. = paste("00:", Mark, sep = "")))
```

```
## Warning in convert.times(times., fmt): NAs introduced by coercion
```

```
## Warning in convert.times(times., fmt): 201 time-of-day entries out of range set
## to NA
```

One more time, since the three remaining events of 400m, 200m, and 100m races all start with the seconds place, we add 00:00: before the Mark to convert it into HH:MM:SS format. While it may appear in the tibble that miliseconds are lost, the values are still kept. The first entry for 2001 (Maurice Greene) was lost for the 100m since we added column names to the dataset, replacing the first row. Therefore, we add Competitor Maurice Greene back into the dataset in this step.

```
s_first <- filter(results, Distance %in% c("400m", "200m", "100m")) %>%
  rbind(c(0,"M","100m",1,9.82,-0.2,"MauriceGREENE","23Jul1974","USA",1,"CommonwealthStadium,Edmonton(CA
  mutate(Time = chron(times. = paste("00:00:", Mark, sep = "")))
```

Finally, we take all of the parts that we created before and bind them together into one dataset called cleaned_results. The variable sixteen is created since the chron package adds 16 hours to each time function. To make sure that the time values are accurate in graphs, we will use this variable and subtract it to the time. All that it will do is make the axis of the graphs more readable.

```
cleaned_results_with2020 <- rbind(h_first, h_first_half, m_first_half, m_first, s_first) %>%
  mutate(Year = strtoi(substr(Date, 6, 9)), Date = as_date(Date, format = "%d%b%Y"), DOB = as_date(DOB,
  select(Gender, Distance, Year, Rank, Competitor, Time, Venue, Date, DOB, Nat)
sixteen = "16:00:00"
```

Case for removing 2020

In our initial exploration of the dataset, we each individually ran data visualizations for assigned race distances and realized that the data points corresponding to the year 2020 behaved quite unexpectedly. As such, we make the following case for these datapoints to be removed as outliers:

In large, we believe the abnormalities observed in the aforementioned data points can be attributed to the rise of a global pandemic at the time. Covid-19 in 2020 halted all track meets and races. In addition, it was supposed to be an Olympics year for the 2020 Tokyo Olympics and as such, prior to track meets and races being shut down, the marks recorded for 2020 were abnormally high. This can be attributed to competitors not racing as hard in order to prepare their legs for the Olympics and produce their best standing at the Olympic races.

The following is a visual demonstration of how the removal of 2020 data points can help us make more accurate predictions:

```r
resMarathon <- cleaned_results_with2020 %>% filter(Distance == "Marathon") %>%
  filter(Gender == "M") %>%
    mutate(asNumMark = as.numeric(Time))

Marathon_pct <- resMarathon %>%
  group_by(Year) %>%
  summarise(mean_time = mean(asNumMark)) %>%
  mutate(pct_change = ((mean_time) - lag(mean_time))/lag(mean_time)*100) %>%
  arrange(mean_time) %>%
  mutate(Event = "Marathon")

res200 <- cleaned_results_with2020 %>% filter(Distance == "200m") %>%
  filter(Gender == "M") %>%
    mutate(asNumMark = as.numeric(Time))

X200_pct <- res200 %>%
  group_by(Year) %>%
  summarise(mean_time = mean(asNumMark)) %>%
  mutate(pct_change = ((mean_time) - lag(mean_time))/lag(mean_time)*100) %>%
  arrange(mean_time) %>%
  mutate(Event = "200m")
```

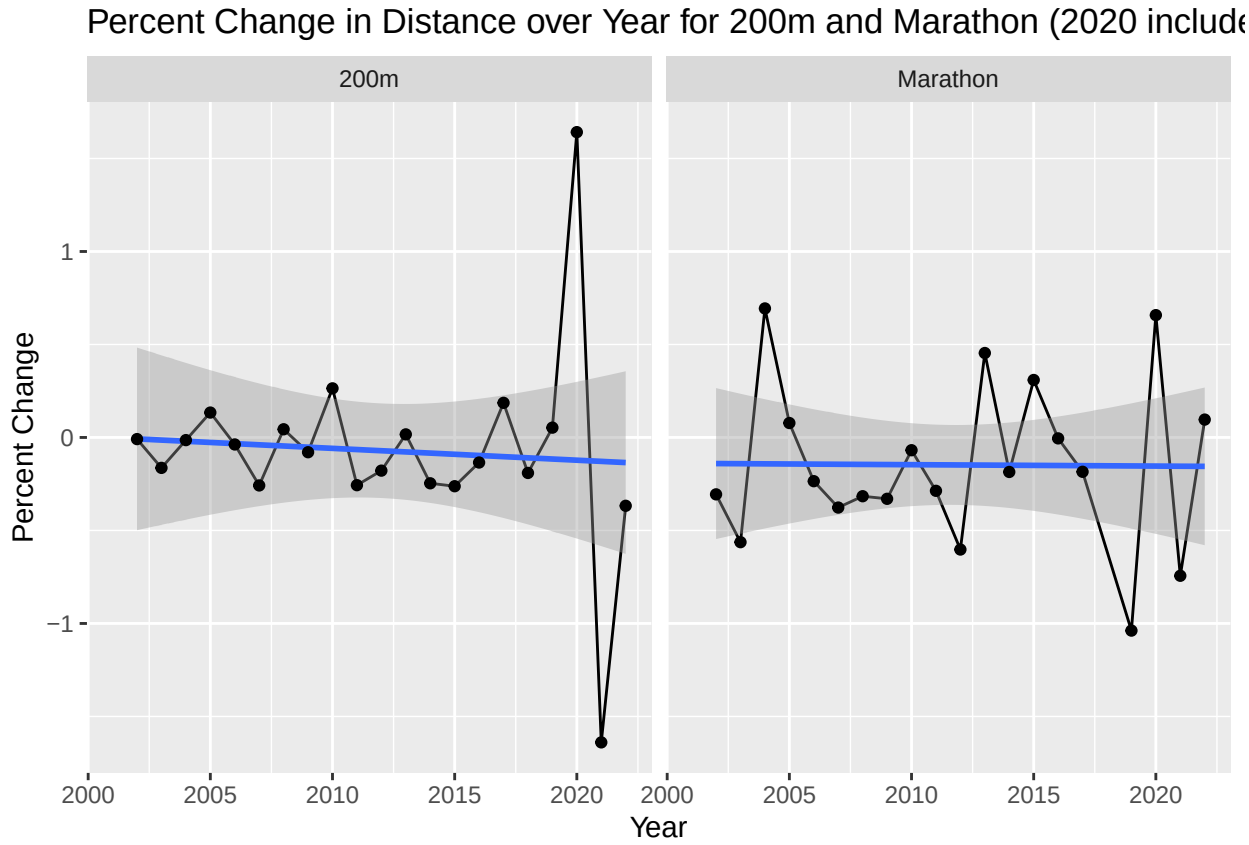
```r
X200_pct %>%
  rbind(Marathon_pct) %>%
  ggplot(aes(x=Year, y=pct_change)) +
  geom_line() +
  geom_smooth(method = "lm") +
  facet_wrap(~Event) +
  geom_point() +
  ylab("Percent Change") +
  ggtitle("Percent Change in Distance over Year for 200m and Marathon (2020 included)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



As shown with the 200m and Marathon percent change graphs, there are large spikes for the year 2020 which will skew the data.

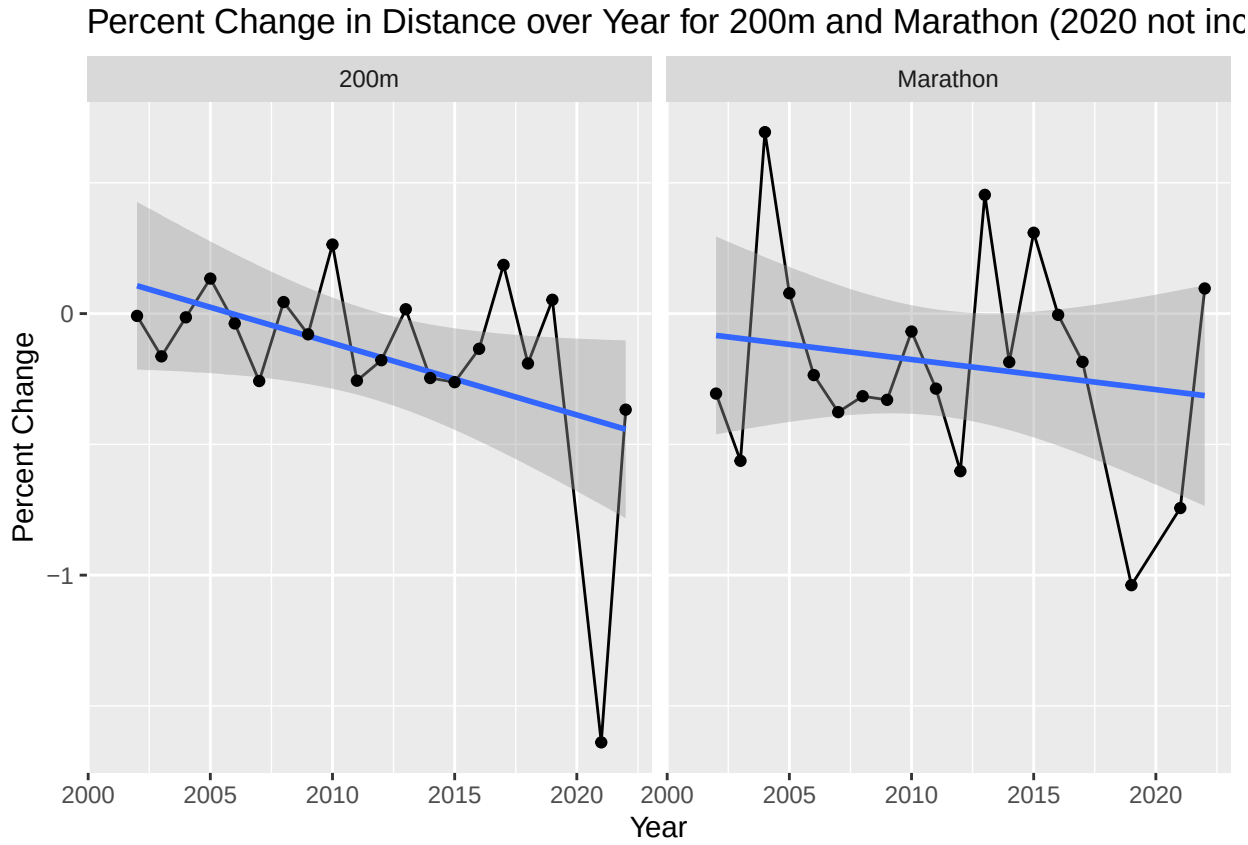
```
X200_pct %>%
  rbind(Marathon_pct) %>%
  filter(Year != 2020) %>%
  ggplot(aes(x=Year, y=pct_change)) +
  geom_line() +
  geom_smooth(method = "lm") +
  facet_wrap(~Event) +
  geom_point() +
  ylab("Percent Change") +
  ggtitle("Percent Change in Distance over Year for 200m and Marathon (2020 not included)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



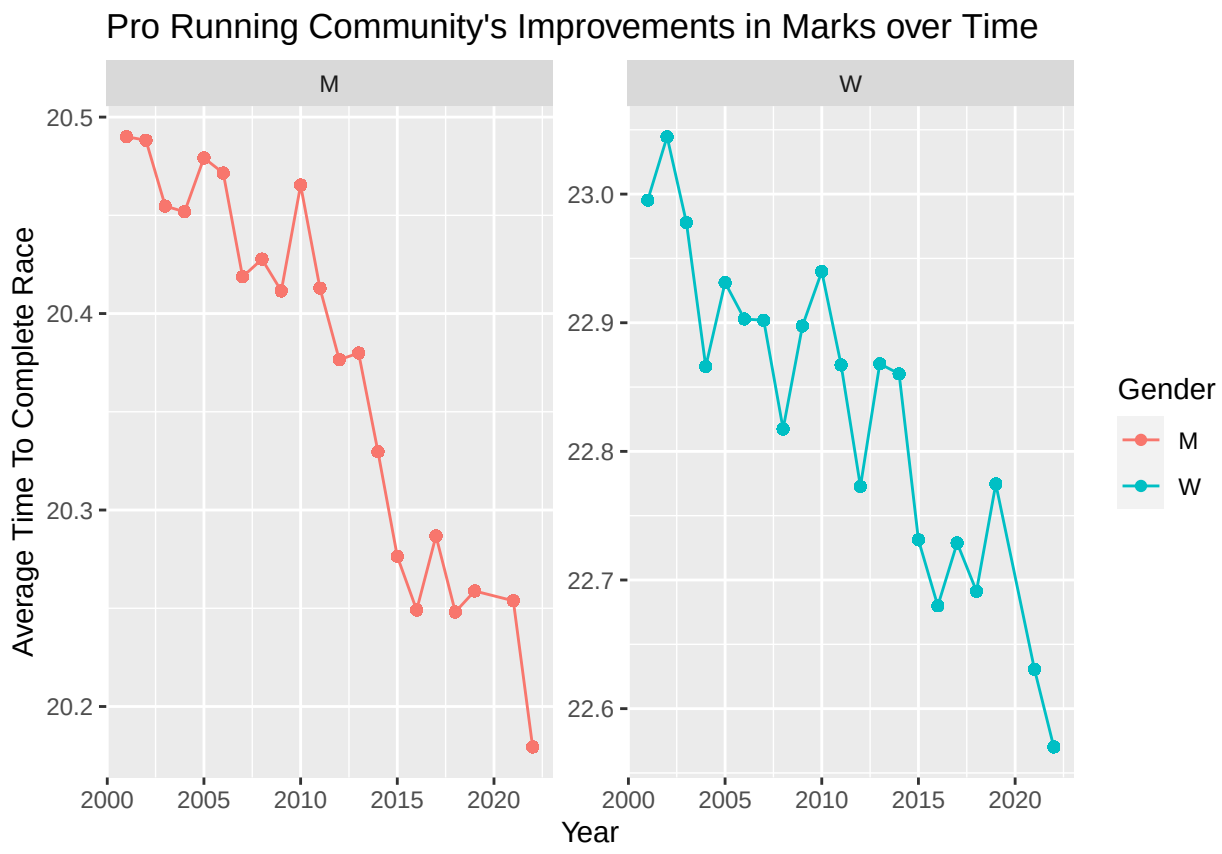
Filtering out 2020 from our data drastically changes the regression models of the graphs meaning that 2020 is a high influence outlier. Therefore, we have decided to remove the year 2020 from our data.

```
cleaned_results <- cleaned_results_with2020%>%
  filter(Year != 2020)
```

Case for choosing men's races only

```
Man200m <- results %>%
  filter(Distance == "200m", Gender == "M") %>%
  mutate(Year = strtoi(substr(Date,6,9))) %>%
  filter(Year !=2020) %>%
  group_by(Year) %>%
  mutate(mean_time = mean(as.numeric(Mark))) %>%
  arrange(mean_time) %>%
  mutate(Dist = "200m")
Woman200m <- results %>%
  filter(Distance == "200m", Gender == "W") %>%
  mutate(Year = strtoi(substr(Date,6,9))) %>%
  filter(Year !=2020) %>%
  group_by(Year) %>%
  mutate(mean_time = mean(as.numeric(Mark))) %>%
  arrange(mean_time) %>%
  mutate(Dist = "200m")
```

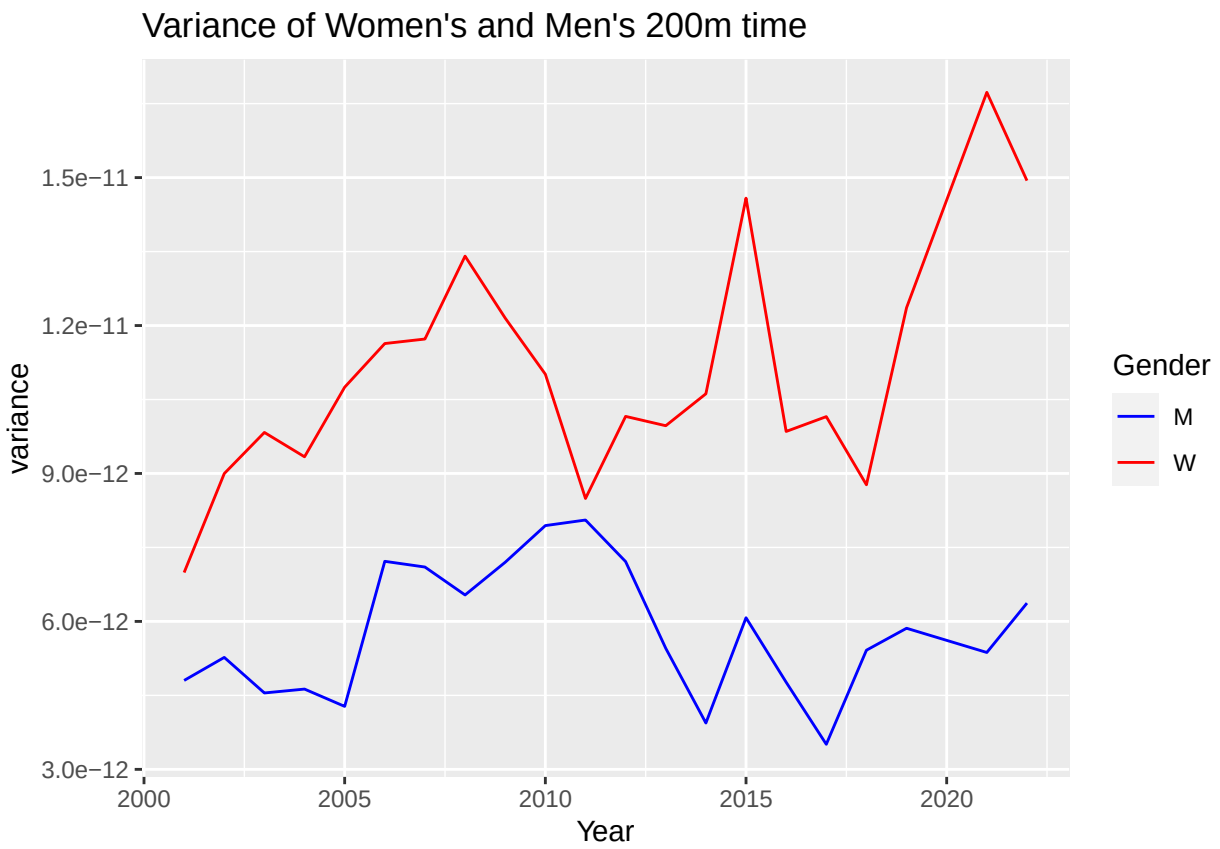
```
genderGraph <- rbind(Man200m, Woman200m)

ggplot(genderGraph, aes(x = Year, y = mean_time, color = Gender)) +
  geom_point() +
  geom_line() +
  facet_wrap(~Gender, scale = "free") +
  labs(x = "Year", y = "Average Time To Complete Race", title = "Pro Running Community's Improvements i
```



We have chosen to stick to analyzing Men's races since in many countries Women's sports is less widely accepted. Therefore the women athletes rarely train unless it is for the olympics. This leads to more variance and larger dips and spikes in the women's races compared to the men's ones. This theory holds true since in the mean times graph for the women's 200m race, the large dips happen every four years, the years when the olympics are held.

```
cleaned_results %>%
  filter(Distance=="200m") %>%
  group_by(Year, Gender) %>%
  summarize(variance=var(Time)) %>%
  arrange(Gender) %>%
  ggplot(aes(x=Year, y=variance, color=Gender)) +
  geom_line() +
  scale_color_manual(values =c("M"="Blue", "W"="Red")) +
  ggtitle("Variance of Women's and Men's 200m time")
```

```
## `summarise()` has grouped output by 'Year'. You can override using the
## `.groups` argument.
```



As shown in the graph provided above, the variance for the women events are larger than that for the mens. The spikes are also more pronounced. As a result of this larger variance and more frequent dips and spikes, we have decided to withhold their races since they made it more difficult to see the observable trends.

Further Reasoning

The following comparison of male race times and female race times goes towards our argument as to why each case should be analyzed individually (i.e. analyses should "facet" in a certain sense by gender):

```
gDif <- cleaned_results %>%
  group_by(Year, Gender, Distance) %>%
  summarise(mean_time = mean(Time)) %>%
  arrange(mean_time)
```

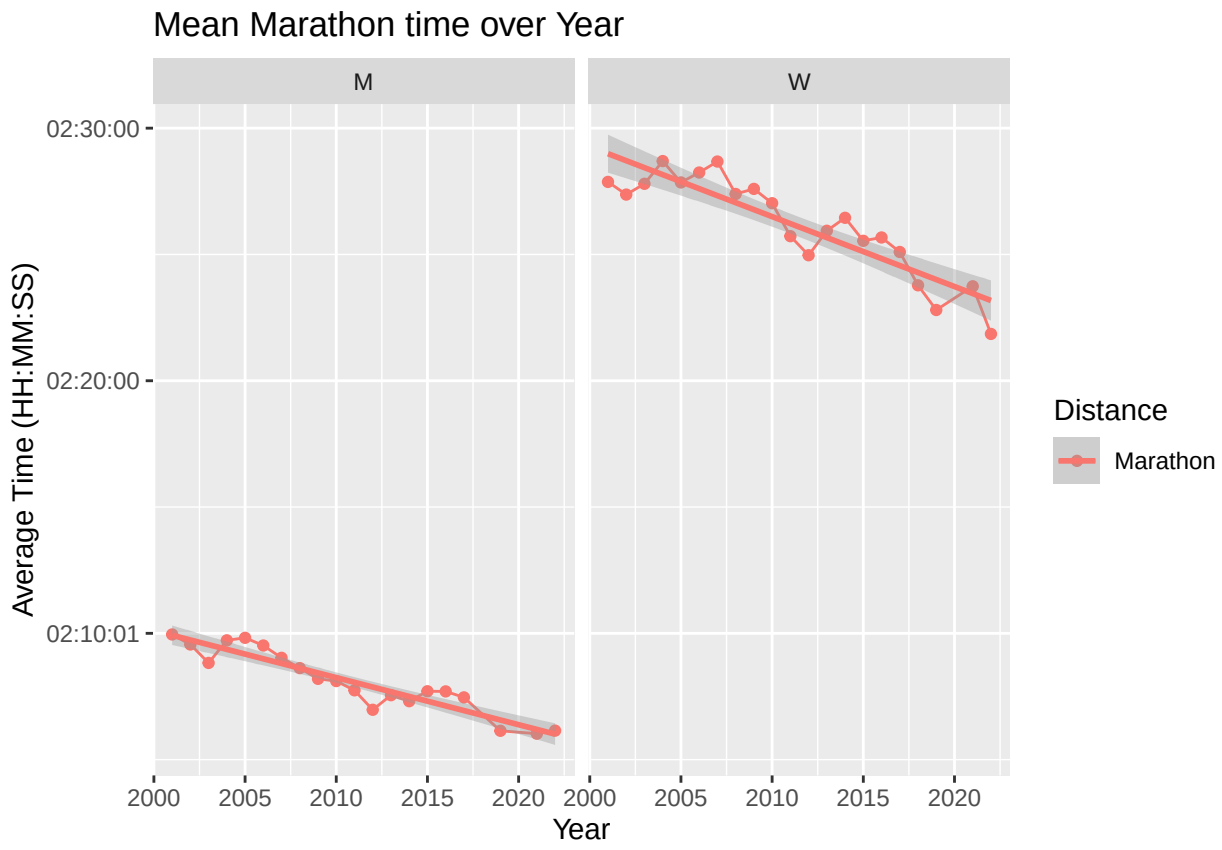
```
## `summarise()` has grouped output by 'Year', 'Gender'. You can override using
## the `.groups` argument.
```

```
gDifMar <- gDif %>% filter(Distance == "Marathon")
gDifHalfMar <- gDif %>% filter(Distance == "Half Marathon")
gDifML <- gDif %>% filter(Distance %in% c("5000m", "1000m"))
gDifMid <- gDif %>% filter(Distance %in% c("400m", "800m"))
gDifSprint <- gDif %>% filter(Distance %in% c("100m", "200m", "400m"))
```

Marathon Comparison

```
ggplot(gDifMar, aes(x=Year, y=mean_time - sixteen, color = Distance)) +
  geom_point() +
  geom_line() +
  geom_smooth(method = "lm") +
  scale_y_chron(format="%H:%M:%S") +
  ylab("Average Time (HH:MM:SS)") +
  ggtitle("Mean Marathon time over Year") +
  facet_wrap(~ Gender)
```

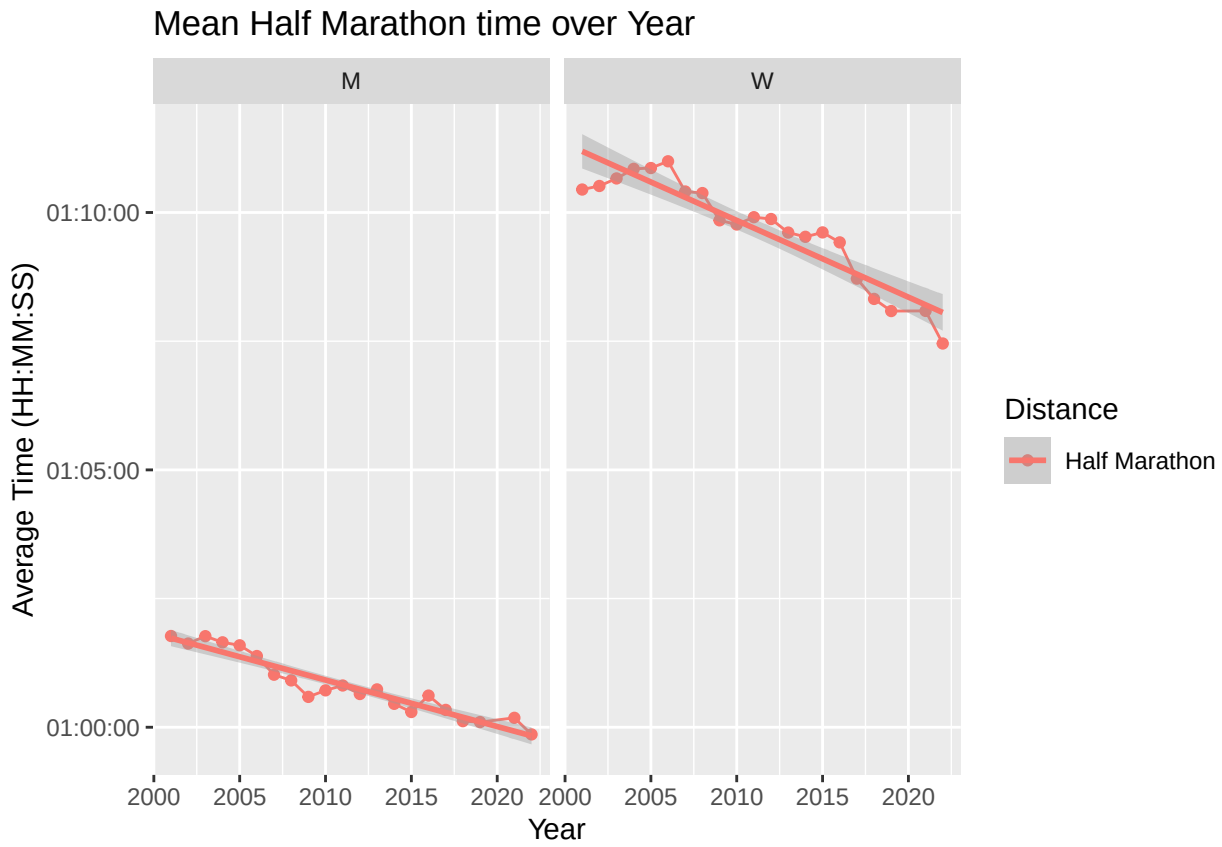
```
## `geom_smooth()` using formula 'y ~ x'
```



Half Marathon Comparison

```
ggplot(gDifHalfMar, aes(x=Year, y=mean_time - sixteen, color = Distance)) +
  geom_point() +
  geom_line() +
  geom_smooth(method = "lm") +
  scale_y_chron(format="%H:%M:%S") +
  ylab("Average Time (HH:MM:SS)") +
  ggtitle("Mean Half Marathon time over Year") +
  facet_wrap(~ Gender)
```

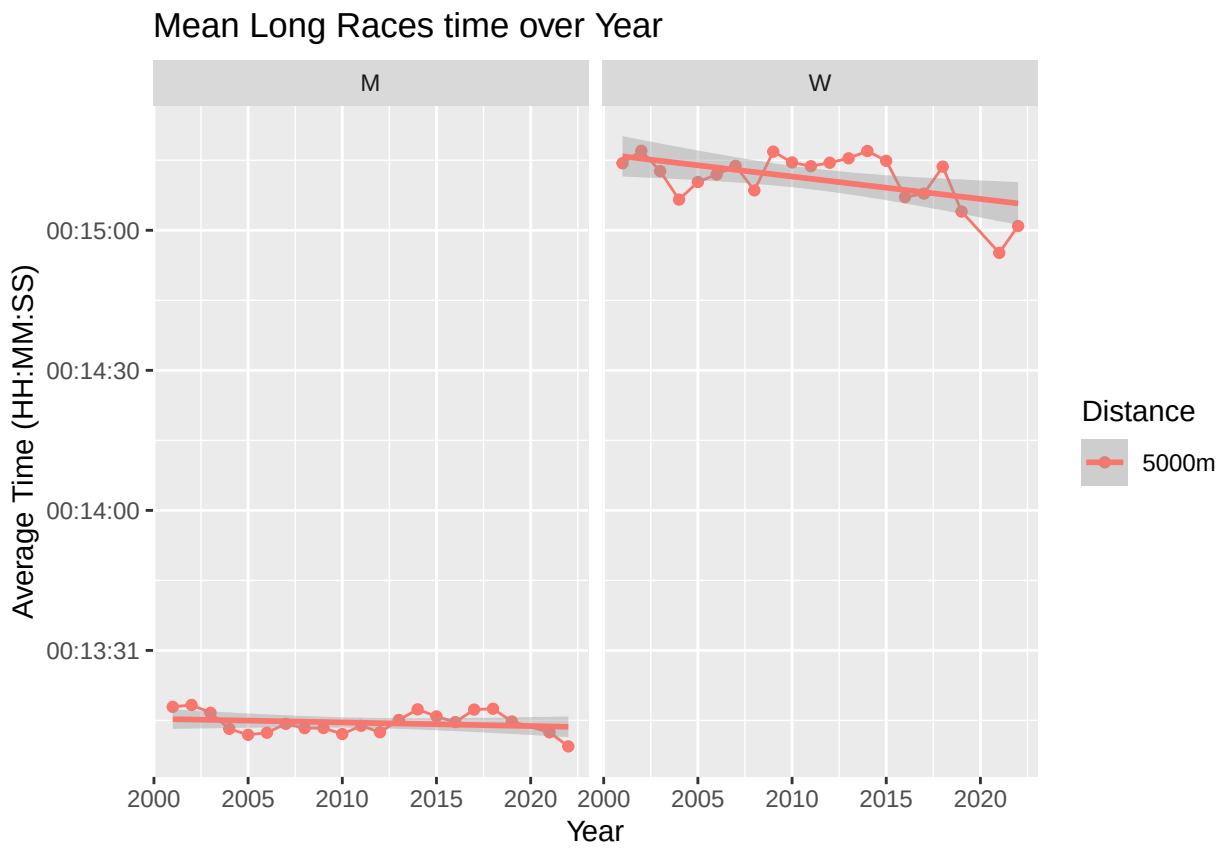
```
## `geom_smooth()` using formula 'y ~ x'
```



Long Races Comparison

```
ggplot(gDifML, aes(x=Year, y=mean_time - sixteen, color = Distance)) +
  geom_point() +
  geom_line() +
  geom_smooth(method = "lm") +
  scale_y_chron(format="%H:%M:%S") +
  ylab("Average Time (HH:MM:SS)") +
  ggtitle("Mean Long Races time over Year") +
  facet_wrap(~ Gender)
```

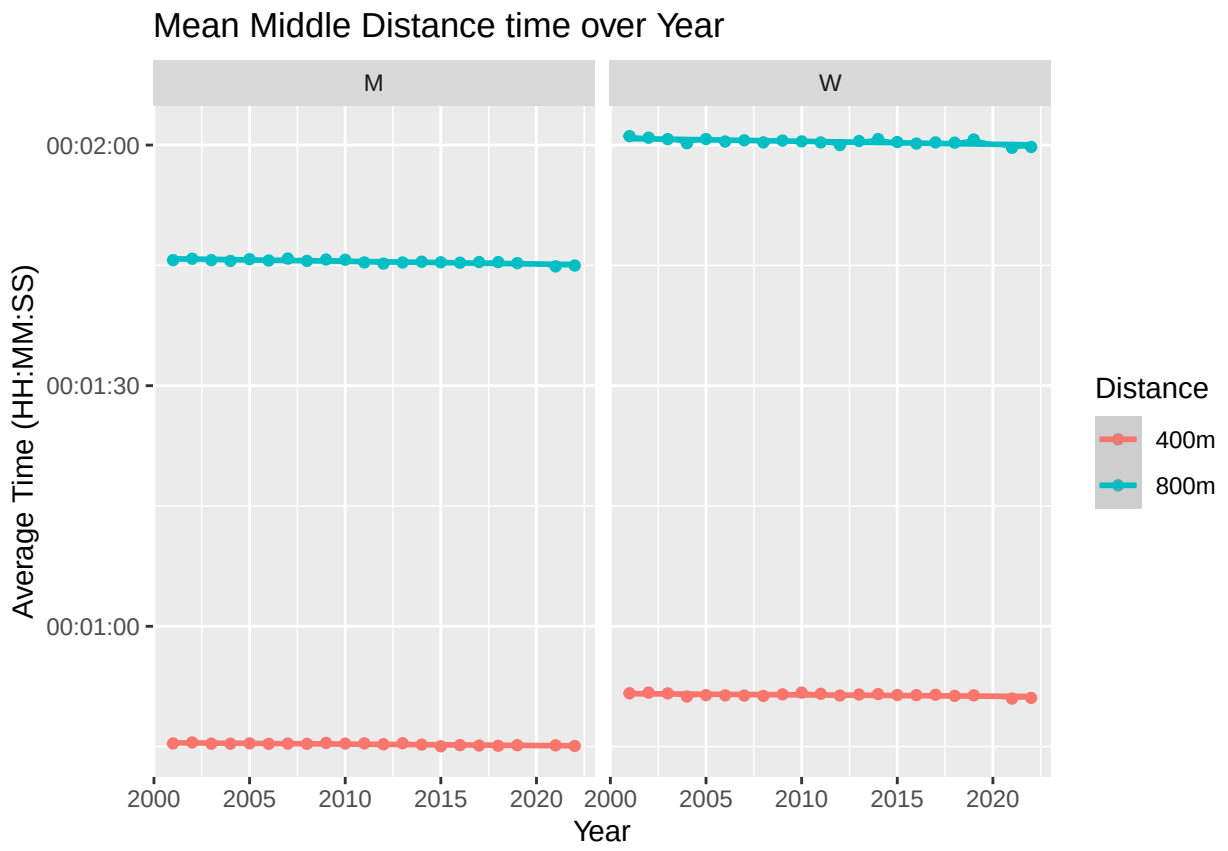
```
## `geom_smooth()` using formula 'y ~ x'
```



Middle Distance Comparison

```
ggplot(gDifMid, aes(x=Year, y=mean_time - sixteen, color = Distance)) +
  geom_point() +
  geom_line() +
  geom_smooth(method = "lm") +
  scale_y_chron(format="%H:%M:%S") +
  ylab("Average Time (HH:MM:SS)") +
  ggtitle("Mean Middle Distance time over Year") +
  facet_wrap(~ Gender)
```

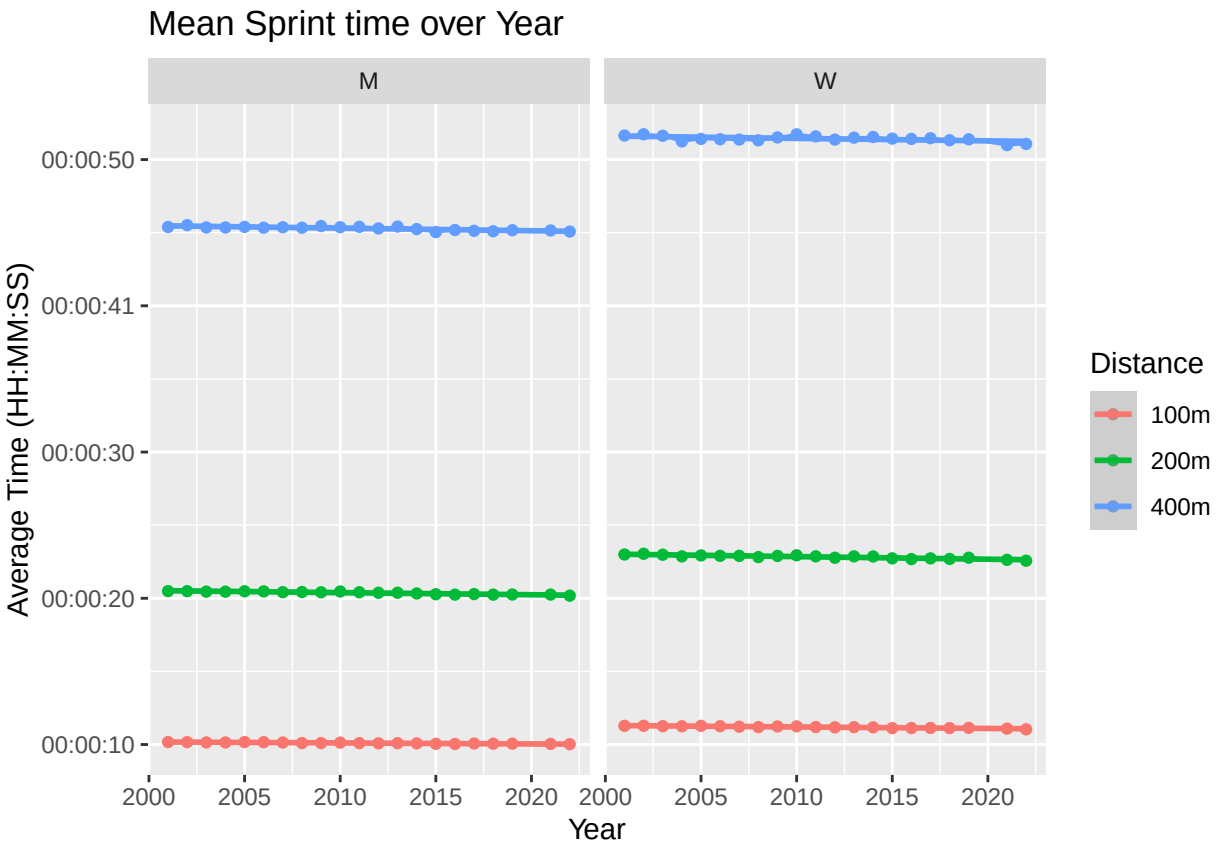
```
## `geom_smooth()` using formula 'y ~ x'
```



Sprint Distance Comparison

```
ggplot(gDifSprint, aes(x=Year, y=mean_time - sixteen, color = Distance)) +
  geom_point() +
  geom_line() +
  geom_smooth(method = "lm") +
  scale_y_chron(format="%H:%M:%S") +
  ylab("Average Time (HH:MM:SS)") +
  ggtitle("Mean Sprint time over Year") +
  facet_wrap(~ Gender)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



The above graphics demonstrate a large difference between male and female race marks, thus proving that this could be a confounding factor. In the interest of producing statistically sound results, we will conduct a considerable portion of our analyses with this in mind. The next 5 code chunks work to filter our data set by gender.

Variable assignment

Marathon

```
Marathon <- cleaned_results %>%
  filter(Gender == "M", Distance == "Marathon")
```

Half Marathon

```
Half_Marathon <- cleaned_results %>%
  filter(Gender == "M", Distance == "Half Marathon")
```

800 meter

```
m800 <- cleaned_results %>%
  filter(Gender == "M", Distance == "800m")
```

200 meter

```
m200 <- cleaned_results %>%
  filter(Gender == "M", Distance == "200m")
```

100 meter

```
m100 <- cleaned_results %>%
  filter(Gender == "M", Distance == "100m")
```

Note:

There are multiple sections where there are values that are removed from the graphs. These warning are coming from two main locations. The first type is when the times are hand counted and lead to variance and inaccuracy. These values should be removed. The second type is when we create the age for the competitors. Some racers only have their year of birth which will not create accurate age's for racers, therefore we want to remove these values too. For each event we have 2000 data points for each type of race.

Data Analysis

Marathon

Marathon Mean time over year

What does the Marathon times look over the years since 2001 to 2022? Let's look into the mean times.

```
MeanMarathon <- Marathon %>%
  group_by(Year) %>%
  summarise(mean_time = mean(Time)) %>%
  arrange(mean_time)
MeanMarathon %>%
  arrange(desc(Year)) %>%
  head(10)
```

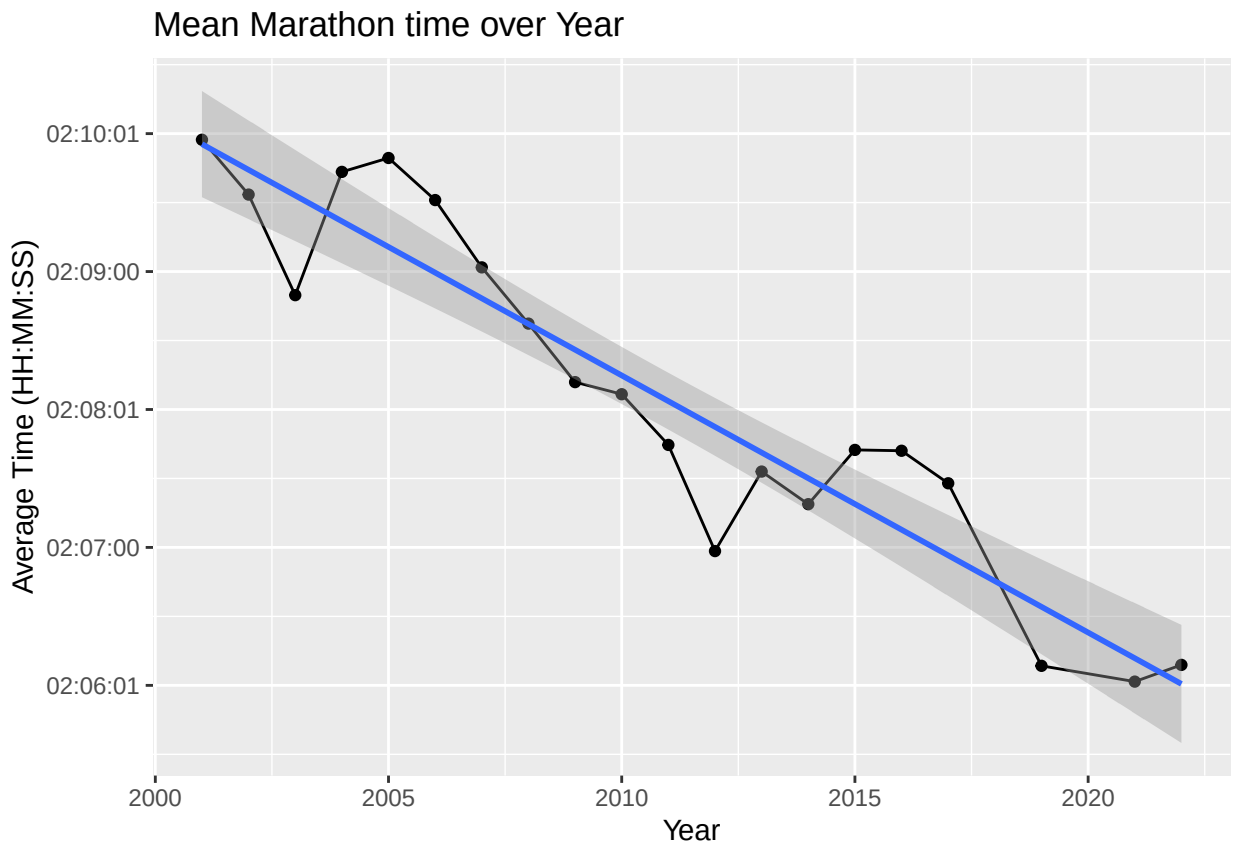
```
## # A tibble: 10 x 2
##     Year mean_time
##    <int> <times>
##  1  2022 02:06:09
##  2  2021 02:06:02
##  3  2019 02:06:08
##  4  2017 02:07:28
##  5  2016 02:07:42
##  6  2015 02:07:42
##  7  2014 02:07:19
##  8  2013 02:07:33
##  9  2012 02:06:58
## 10  2011 02:07:45
```

The values show that the mean times for Marathon runners have consistently decreased throughout the years, but there is a major dip in the times in 2019 that stayed consistent after. Exploring the context of these values, we see that the times began to drastically dip right around Nike's release of the new revolutionary Vaporfly

Next% running shoes. This shoe was more widely adopted by the runners due to its advanced technology and wider base that provided more stability for runners that did not have perfect running mechanics.

```
ggplot(MeanMarathon, aes(x=Year, y=mean_time - sixteen)) +
  geom_point() +
  geom_line() +
  geom_smooth(method = "lm") +
  scale_y_chron(format="%H:%M:%S") +
  ylab("Average Time (HH:MM:SS)") +
  ggtitle("Mean Marathon time over Year")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



There are large dips in the time between 2017 and 2019. This significant decrease is likely attributed to the Vaporfly Next% which was more widely accepted than the Vaporfly 4% that was developed for Eulid Kipchoge by Nike. This shoe did not require perfect running mechanics to feel stable so runners trust it more and allowed the racers to feel more comfortable running at faster speeds. When a shoe is not stable, it leads to runners trusting their gear less and not going at their top performance.

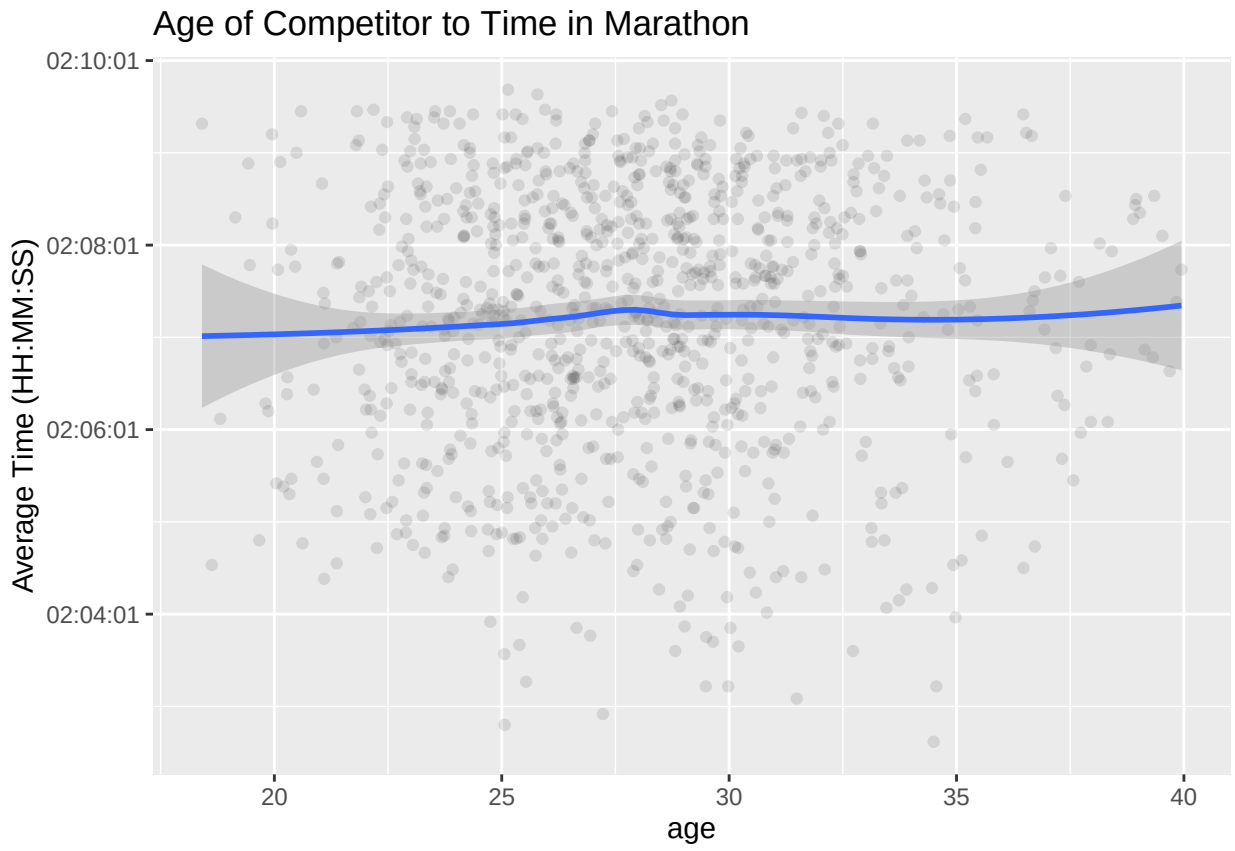
Marathon Age over time mark

Can we attribute this change to age of runners? Marathon is a race where people age into as they gain experience pacing and running. Is age a factor? 2019 saw the introduction of many new, younger racers (Lelisa Desisa and Joyciline Jepkosgei) . Could this explain major time dips in 2019 and 2017?

```
MarathonAge <- Marathon %>%
  mutate(age = (Date-DOB)/365) %>%
  filter(Rank > 10) %>%
  filter(year(Date) >=2010)
MarathonAge %>%
  filter(age <= 40) %>%
  ggplot(aes(x=age, y=Time-sixteen)) +
  geom_point(alpha = 0.1) +
  geom_smooth(method = "loess") +
  scale_y_chron(format="%H:%M:%S") +
  ylab("Average Time (HH:MM:SS)") +
  ggtitle("Age of Competitor to Time in Marathon")
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```

```
## `geom_smooth()` using formula 'y ~ x'
```



We have decided to filter out age for after 2010 because for both long distance road races and sprints, technological improvements mainly happen in 2016. So we filter out for years after 2010 to center the technological improvements to see how we have the same number of data points before and after technology was introduced to get similar comparisons.

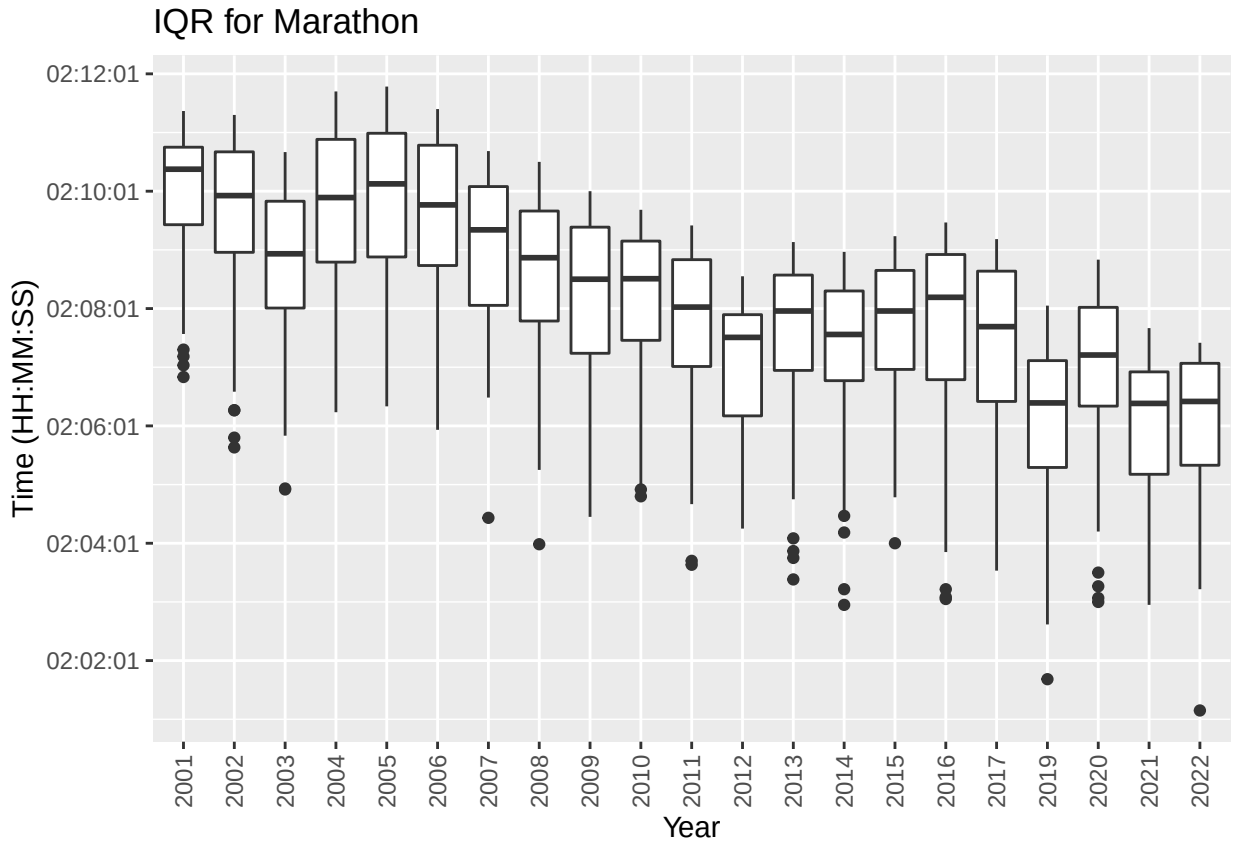
While being younger seems to be correlated with an initial advantage (shown by the first dip in times), after that, increased age slowly makes times trend downwards, implying that increased training leads to more efficient form, emphasizing how important efficiency is in the event.

At the younger ages, athletes are more explosive, but less efficient, something that is important in marathon running. In addition, younger runners are less experienced so they have more difficulty pacing themselves and understanding the strategy that goes into marathon running. At the older ages, the runners are less explosive so they are not able to maintain the top ideal speed for long enough. However, they are still able to place in the top 100 results for any given year because of their experience running and greater understanding of the marathon event.

Marathon IQR graph

How about we look into the distrubtion of values in these years? Let's create a boxplot IQR graph for these times.

```
results %>%
  filter(Distance == "Marathon", Gender == "M") %>%
  mutate(Year = substr(Date, 6, 9), Time = chron(times. = Mark)) %>%
  select(Competitor, Time, Year) %>%
  ggplot(aes(x=Year, y=Time - sixteen)) + geom_boxplot() + scale_y_chron(format="%H:%M:%S") +
  ylab("Time (HH:MM:SS)") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  ggtitle("IQR for Marathon")
```




```
results %>%
  filter(Distance == "Marathon", Gender == "M") %>%
  mutate(Year = substr(Date, 6, 9), Time = chron(times. = Mark)) %>%
```

```
  select(Competitor, Time, Year) %>%
  group_by(Year) %>%
  summarize(median_time = median(Time), q1 = quantile(Time, 0.25)) %>%
  filter(Year %in% c(2016,2017,2019,2020))
```

```
## # A tibble: 4 x 3
##   Year  median_time q1
##   <chr> <times>     <times>
## 1 2016  02:08:12    02:06:47
## 2 2017  02:07:42    02:06:25
## 3 2019  02:06:24    02:05:18
## 4 2020  02:07:12    02:06:20
```

As shown, the median time for the Marathon in 2019 was faster than the Quartile 1 time in 2017, meaning that as a collective, the entire running scene was getting faster and widely adopted the Nike Vaporfly marathon shoes.

Nike's Breaking2 project that was designed to get an athlete to run the marathon below 2 hours started in 2016 and led to large amounts of research and development for the company in everything from shoe technology to diet. The Nike sponsored athletes had exclusive access to this technology that led to significant outliers in 2016. In 2017, the Nike Vaporfly 4% which used to be exclusive to Nike athletes were released to the general public which is why the outliers are no longer existent.

2017 saw a slight improvement in the median and Q3 times, but other quartiles were still behind as people did not fully accept that the 4% was better due to its unprecedented design and stability issues. People in the past thought that a shoe that was lower to the ground and thinner would be better due to the increased road feel and decreased weight. In contrast, the Nike Vaporfly 4% were high to the ground and had a large midsole.

In 2019, the Nike Vaporfly Next% were released and featured a wider base that improved the stability of the shoe. This lead to more people adopting the shoe and led to the whole population of marathon runners to improve. The median in 2019 was now faster than the top 25% of 2017. The median in 2019 was faster than the top 25% time in 2017 by 1 second as shown in the table.

In the years after 2019, there have not been major shifts in the distribution of the quartiles showing that technology in marathon running has not improved significantly since then. The other companies such as Adidas have tried to imitate Nike's shoe by using similar foam and a carbon plate for structure.

Let's look into who the outliers were for the years 2016, 2019, and 2022. Who are these people and why were they so fast?

```
Marathon %>%
  filter(Year==2016) %>%
  arrange(Time) %>%
  select(Competitor, Year, Time) %>%
  head(3)
```

```
##                Competitor Year     Time
## 1         KenenisaBEKELE 2016 02:03:03
## 2          EliudKIPCHOGE 2016 02:03:05
## 3 WilsonKipsangKIPROTICH 2016 02:03:13
```

Two of the outliers seen in 2016 were Nike athletes (Bekele, Kipchoge) and the third athlete (Kiprotich) was found to be doping in 2016. https://www.independent.co.uk/sport/general/athletics/wilson-kipsang-kiprotich-ban-landslide-missed-tests-kenya-a9600416.html

```
Marathon %>%
  filter(Year==2019) %>%
  arrange(Time) %>%
  select(Competitor, Year, Time) %>%
  head(1)
```

```
##       Competitor Year     Time
## 1 KenenisaBEKELE 2019 02:01:41
```

There is one major outlier for 2019, this being Bekele, a Nike sponsored runner. He had longer time with the shoes that the other runners got. He has been using the Nike Vaporfly Next% since 2017

```
Marathon %>%
  filter(Year==2022) %>%
  arrange(Time) %>%
  select(Competitor, Year, Time) %>%
  head(3)
```

```
##      Competitor Year     Time
## 1 EliudKIPCHOGE 2022 02:01:09
## 2   AmosKIPRUTO 2022 02:03:13
## 3   TamiratTOLA 2022 02:04:14
```

2022 has a large outlier since Kipchoge, the athlete selected by Nike to Break2 used a shoe designed for him, this being the Nike Alphafly. Other athletes do not adopt this shoe as it is developed precisely for his form.

What are the shapes of the graphs from 2015-2022? Based on the IQR graphs: * 2015 was a tight IQR with a skew toward the shorter times (left) meaning that natural ability was a large component to a time. * 2016 was a slightly larger IQR and was skewed heavily towards shorter times (left) with three outliers. This is due to the discrepancy of Nike Breaking 2 athletes. * 2017 was an even larger IQR. The Q1 group got faster. The value of Q1 got a lot faster. The shape is still slightly skewed to shorter time (left). Q1 represents the top 25% and the adoption rate of the Nike Vaporfly 4% was likely around 25%-30%. * 2018 (not included as mentioned before) * 2019 all data points heavily move down. The skew is similar to 2015, implying that the technology gap has closed since the majority of racers have adopted the new Nike Vaporfly Next%. * 2021-2022 were somewhat identical to 2019 as technology in shoes has not really changed since then.

T test for Marathon 2017 and Marathon 2019

We are testing the null hypothesis that there is no variation between the years and that the shoes did not provide a statistically significant difference. Is there a statistically significant difference in times between years using the before and after and the Vaporfly Next% (most widely accepted)?

```
Marathon2017 <- cleaned_results %>%
  filter(Distance=="Marathon", Gender=="M", Year==2017)

Marathon2019 <- cleaned_results %>%
  filter(Distance=="Marathon", Gender=="M", Year==2019)

tMarathon <- Marathon2017 %>%
  rbind(Marathon2019)

t.test(Time ~ Year, data = tMarathon)
```

```
##
##  Welch Two Sample t-test
##
## data:  Time by Year
## t = 6.86552, df = 197.96, p-value = 0
## alternative hypothesis: true difference in means between group 2017 and group 2019 is not equal to 0
## 95 percent confidence interval:
##  00:00:57 00:01:42
## sample estimates:
## mean in group 2017 mean in group 2019
##           02:07:28           02:06:08
```

Assuming there is no change in the general downward trend due to advancements in Marathon-specific technology from 2017-2019, the probability of the graph ever taking on a value such as this randomly is astronomically low with a p value of almost 0 (as denoted by the very large t-test).

Linear Regression Model

Since we know that this is statistically significant, how does a runner's 2017 time conver to 2019?

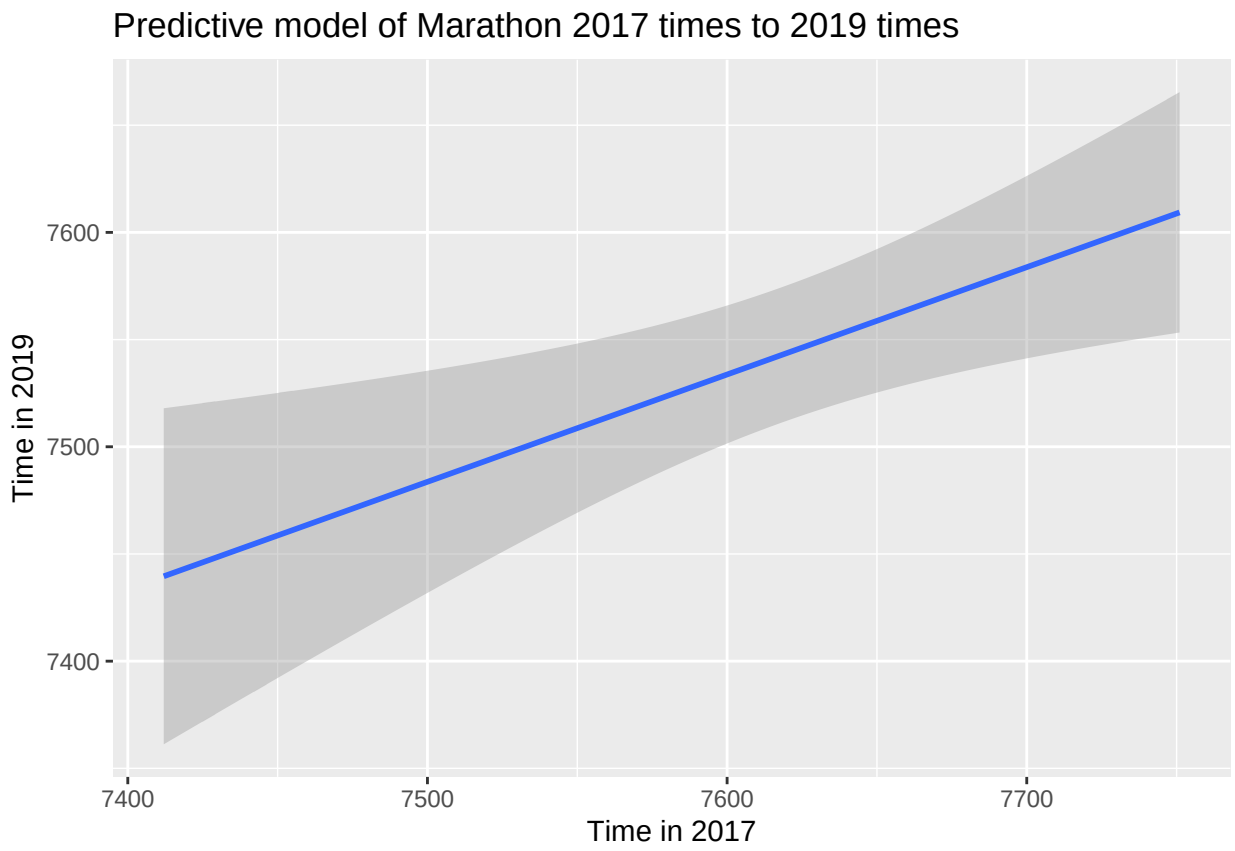
```
Marathon2017 <- cleaned_results %>%
  filter(Distance=="Marathon", Gender=="M", Year==2017) %>%
  mutate(Time = hours(Time)*3600 + minutes(Time)*60 + seconds(Time))

Marathon2019 <- cleaned_results %>%
  filter(Distance=="Marathon", Gender=="M", Year==2019) %>%
  mutate(Time = hours(Time)*3600 + minutes(Time)*60 + seconds(Time))

join_Marathon <- inner_join(Marathon2017, Marathon2019, by="Competitor")
lm(Time.y~Time.x, data= join_Marathon)
```

```
##
## Call:
## lm(formula = Time.y ~ Time.x, data = join_Marathon)
##
## Coefficients:
## (Intercept)       Time.x
##   3728.7665       0.5007
```

We convert the Time into seconds to get more accurate results of this linear regression. The gap of impact of competition reduced in 2019 because the slope of 0.5 predict that for each second faster one was in 2017, they are only a half second faster in 2019

```
join_Marathon %>%
  ggplot(aes(x=Time.x, y=Time.y)) +
  geom_smooth(method="lm") +
  ggtitle("Predictive model of Marathon 2017 times to 2019 times")+
  xlab("Time in 2017") +
  ylab("Time in 2019")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Further research

We tried to create a dataset of what top three finishers wore in the Berlin Marathon. However this data was hand collected and hand recoreded only using the top 3 racers. We wanted to use this data including the shoe specifications to be able to see which specification mattered the most for race improvement

```
Shoes <- read.csv("shoesdata.csv")
Shoes %>%
  head(12)
```

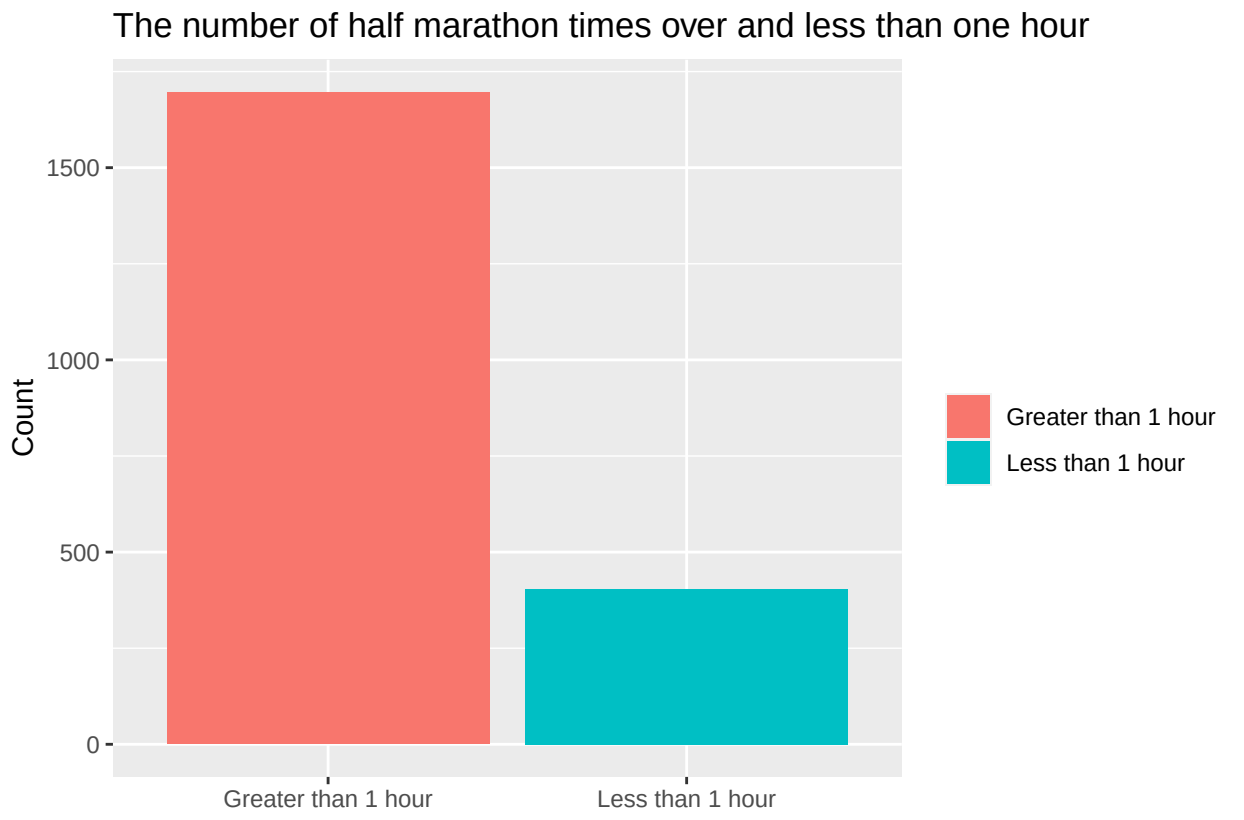
```
##    Distance     Competitor Year                     shoe_name shoe_weight
## 1  Marathon  EliudKIPCHOGE 2022 Nike Air Zoom Alphafly Next% 2        8.40
## 2  Marathon      MarkKORIR 2022     Adidas AdiZero Adios Pro 3        8.40
## 3  Marathon      TaduABATE 2022        Li Nang Feidian 3 Ultra        6.60
## 4  Marathon      GuyeADOLA 2021     Adidas AdiZero Adios Pro 2        7.60
## 5  Marathon   BethwelYEGON 2021       Adidas AdiZero Adios Pro        8.00
## 6  Marathon KenenisaBEKELE 2021          Nike Vaporfly Next% 2        6.90
## 7  Marathon KenenisaBEKELE 2019            Nike Vaporfly Next%        7.10
## 8  Marathon  BirhanuLEGESE 2019            Nike Vaporfly Next%        7.10
## 9  Marathon     SisayLEMMA 2019            Nike Vaporfly Next%        7.10
## 10 Marathon  EliudKIPCHOGE 2018         Nike Vaporfly Elite v2        7.10
## 11 Marathon    AmosKIPRUTO 2018           Adidas AdiZero Adios        8.35
## 12 Marathon  WilsonKIPSANG 2018            Adidas AdiZero Sub2        6.25
##    shoe_drop full_length_cf
## 1        8.0            yes
```

```
## 2        6.5           yes
## 3        7.0            no
## 4       10.0           yes
## 5        8.5           yes
## 6        8.0           yes
## 7        9.0           yes
## 8        9.0           yes
## 9        9.0           yes
## 10       9.0           yes
## 11       9.5           yes
## 12       6.0            no
```

Half Marathon

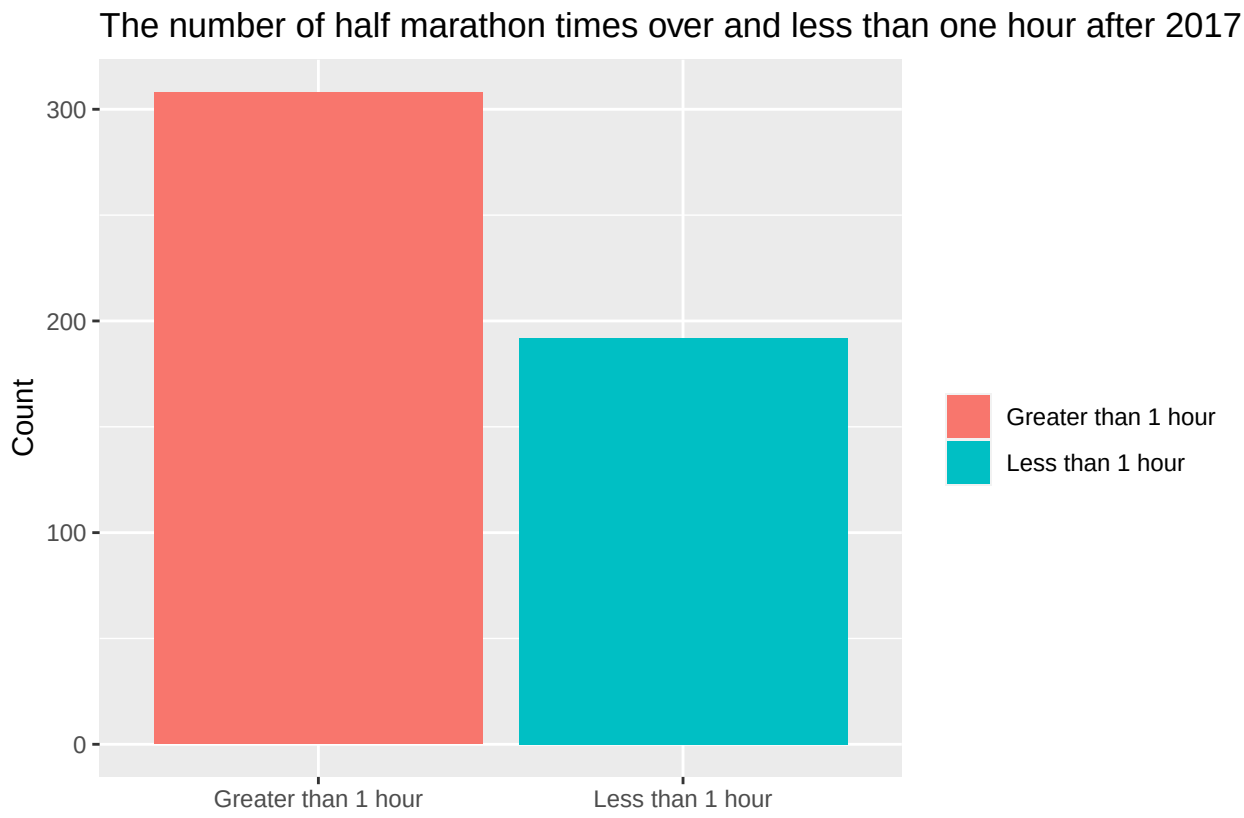
As our analysis thus far has been largely marathon-centered, the next few pages will focus on half-marathons. Half Marathons normally come in around the 1 hour mark. So how many of the top 100 times in years since 2001 are under 1 hour? First, we analyze the number of runners that finished their half-marathons under the 1 hour mark vs the amount that finished over. This gives us an initial understanding to compare future graphs to:

```
LongerHalf_Marathon <- Half_Marathon %>%
  filter(Time >= "01:00:00") %>%
  mutate(longer_hour = "Greater than 1 hour")
ShorterHalf_Marathon <- Half_Marathon %>%
  filter(Time < "01:00:00") %>%
  mutate(longer_hour = "Less than 1 hour")

HalfMarathonDistribution <- ShorterHalf_Marathon %>%
  rbind(LongerHalf_Marathon) %>%
  group_by(longer_hour)

HalfMarathonDistribution %>%
  count() %>%
  ggplot(aes(x=longer_hour, y=n, fill = longer_hour)) +
  geom_col() +
  xlab("") +
  ylab("Count") +
  ggtitle("The number of half marathon times over and less than one hour") +
  guides(fill=guide_legend(title=""))
```



As can be seen, most half Marathon times are still longer than that 1 hour mark, however, there have been some times that have started to crack through that barrier. Lets look into more recent years. To see if technology has made a significant impact on half-marathon-runners marks, we will compare this graphic with one that zooms in on half-marathons after 2017. We expect to see the two bars be closer in height, indicating that a higher proportion of runners are breaking the 1 hour mark due to further advancing running shoe technology.

```
HalfMarathonDistribution %>%
  filter(Year >= 2017) %>%
  count() %>%
  ggplot(aes(x=longer_hour, y=n, fill = longer_hour)) +
  geom_col() +
  xlab("") +
  ylab("Count") +
  ggtitle("The number of half marathon times over and less than one hour after 2017") +
  guides(fill=guide_legend(title=""))
```



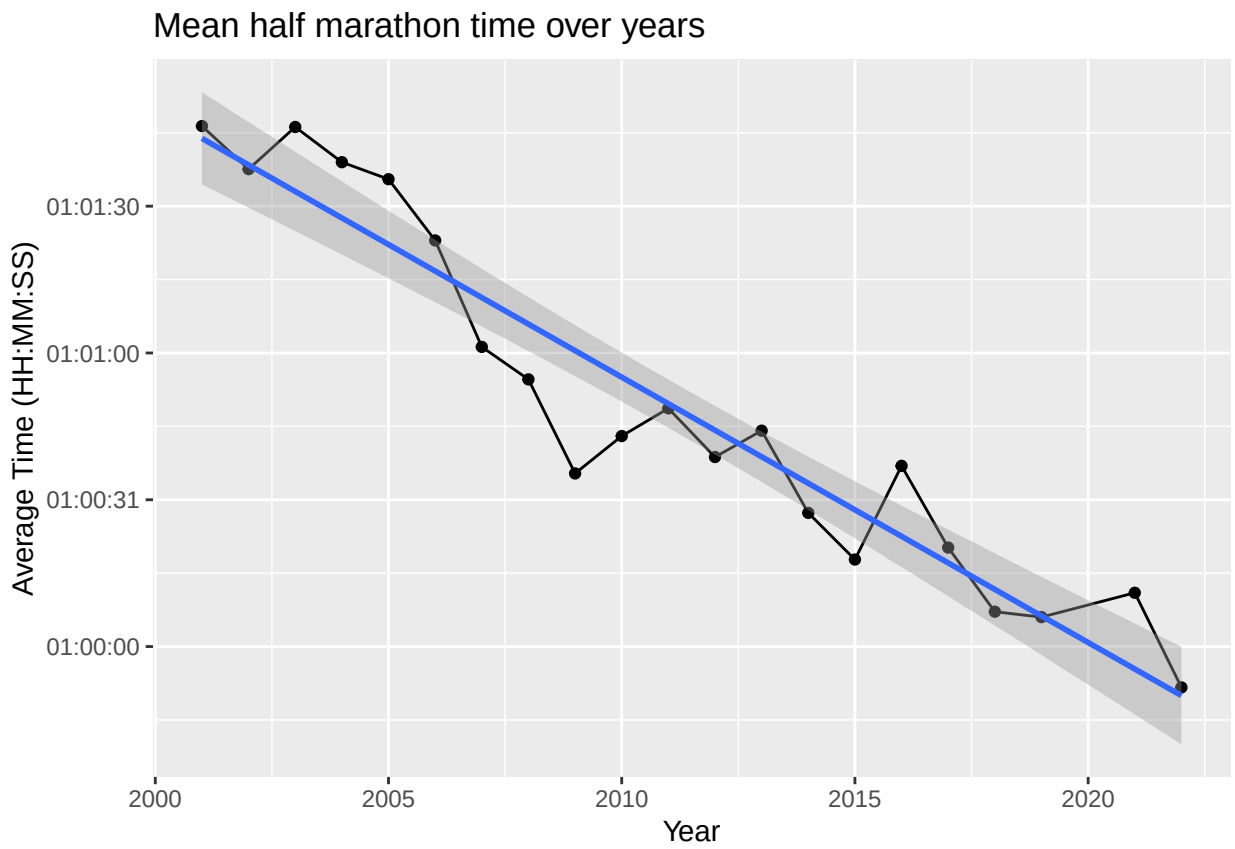
The graphic above serves as evidence for our previous prediction. To confirm further this concept, we next looked at half-marathon mean finish times by year to see when significant changes took place:

Half Marathon Mean time

What do the mean times for half marathon look like? Let's look into the mean times over the years.

```
MeanHalf_Marathon <- Half_Marathon %>%
  group_by(Year) %>%
  summarise(mean_time = mean(Time)) %>%
  arrange(mean_time)
ggplot(MeanHalf_Marathon, aes(x=Year, y=mean_time - sixteen)) +
  geom_point() +
  geom_line() +
  geom_smooth(method = "lm") +
  scale_y_chron(format="%H:%M:%S") +
  ylab("Average Time (HH:MM:SS)") +
  ggtitle("Mean half marathon time over years")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



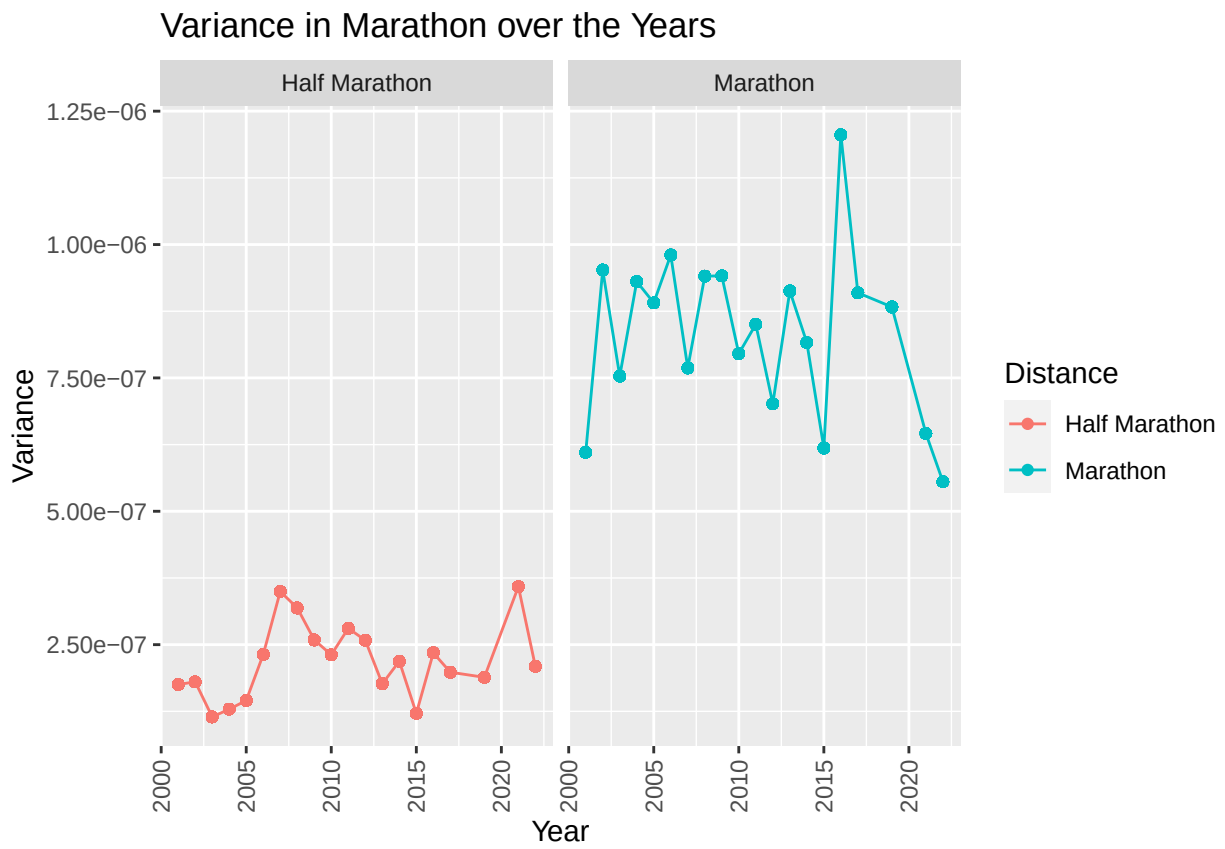
Similar to that seen in the marathon graph, there are large improvements in times in 2017 due to the release of the Vaporfly4%. However, there is a larger drop in Half Marathon time in 2017 than that for the Marathon. Why is this the case?

Variance for Half and Full Marathon

If only certain atheletes had certain shoes in 2016 and people decided not to switch in 2017, how does the spread and accesiblity of technology changes?

```
cleaned_results %>%
  filter(Distance %in% c("Marathon", "Half Marathon"), Gender == "M") %>%
  filter(Year != 2018) %>%
  group_by(Year, Distance) %>%
  summarize(variance = var(Time), Distance) %>%
  filter(Year >= 2001) %>%
  ggplot(aes(x=Year, y=variance, color=Distance)) +
  geom_point() +
  geom_line() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  ylab("Variance") +
  facet_wrap(~Distance) +
  ggtitle("Variance in Marathon over the Years")
```

```
## `summarise()` has grouped output by 'Year', 'Distance'. You can override using
## the `.groups` argument.
```



For marathon, the variance in 2016 was around 1.2* 10^-6, however the variance in 2016 for half marathon is around 2.4*10^-7. There is less variance for the half marathon suggesting the ideas that we have discussed before regarding the athletes being more likely to adopt new technology. The reason for the large variance in Marathon races was because of runners did not want to use the new shoes and runners not having access to the Nike shoes.

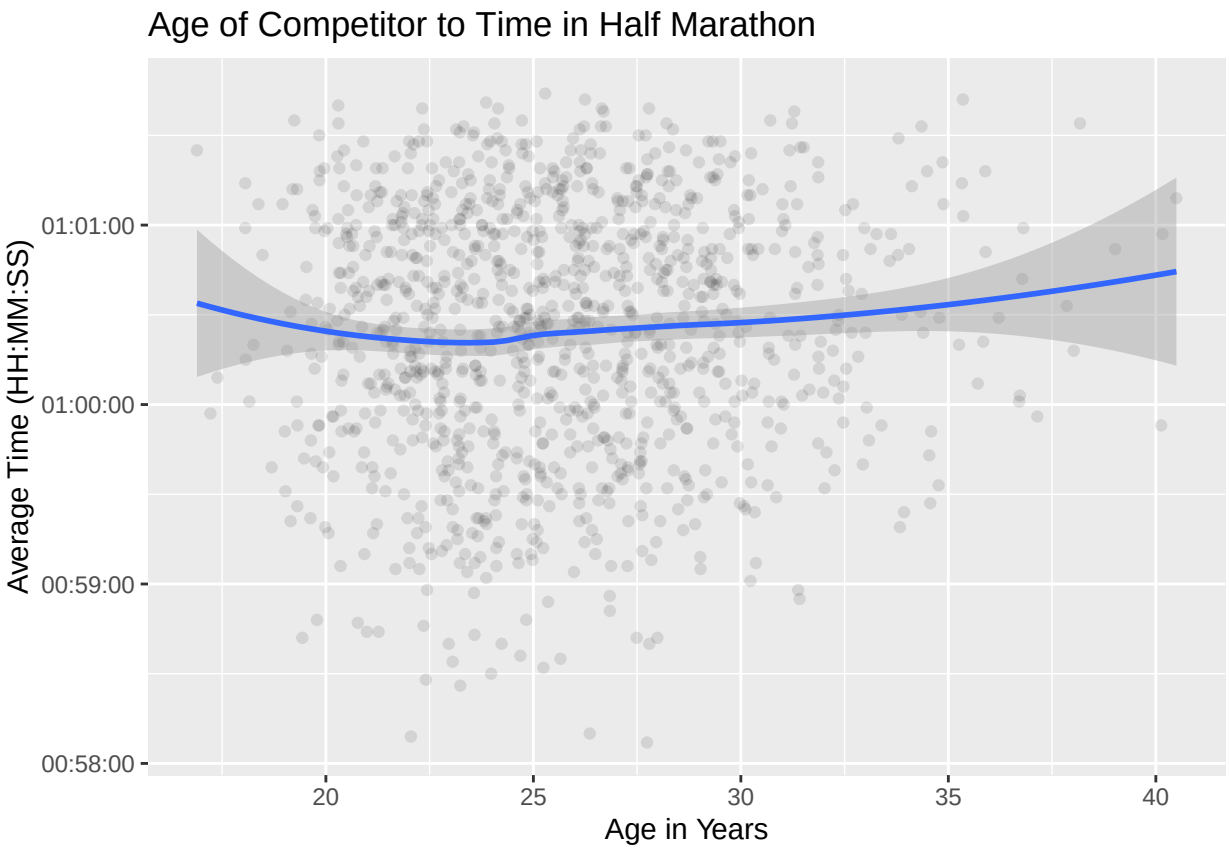
Half Marathon Age over time mark

So why is it less prestigious of a race?

```
Half_MarathonAge<- Half_Marathon %>%
  mutate(age = (Date-DOB)/365) %>%
  filter(Rank > 10) %>%
  filter(year(Date) >=2010)

Half_MarathonAge %>%
  ggplot(aes(x=age, y=Time-sixteen)) +
  geom_point(alpha = 0.1) +
  geom_smooth(method = "loess") +
  scale_y_chron(format="%H:%M:%S") +
  ylab("Average Time (HH:MM:SS)") +
  xlab("Age in Years") +
  ggtitle("Age of Competitor to Time in Half Marathon")
```

```
## Don't know how to automatically pick scale for object of type difftime. Defaulting to continuous.
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 63 rows containing non-finite values (stat_smooth).

## Warning: Removed 63 rows containing missing values (geom_point).
```



The age of half marathon runners is less than that of marathon runners due to the transition that runners often take to reach the marathon.

5 Number Summary of Marathon Age

```
MarathonAge %>%
  mutate(age = as.numeric(age*365, units="days")) %>%
  mutate(age= age/365) %>%
  select(age) %>%
  drop_na() %>%
  summary()
```

```
##       age
##  Min.   :18.41
##  1st Qu.:25.05
##  Median :27.89
##  Mean   :28.06
##  3rd Qu.:30.52
##  Max.   :43.85
```

5 Number Summary of Half Marathon Age

```
Half_MarathonAge %>%
  mutate(age = as.numeric(age*365, units="days")) %>%
  mutate(age= age/365) %>%
  select(age) %>%
  drop_na() %>%
  summary()
```

```
##       age
##  Min.   :16.89
##  1st Qu.:22.76
##  Median :25.22
##  Mean   :25.66
##  3rd Qu.:28.10
##  Max.   :40.50
```

The median for marathon runners is around 28 years while for half marathon runners it is around 25.25 years. This is in large due to the fact that younger runners looking to race marathoan start with half the distance to understand how pacing and efficiency works in long distance road races. At every quartile, the age for half marathon runners is a couple of years less than that of the marathon runners suggesting that the half marathon is a stepping tool for the marathon, otherwise, we would see that older runners would stay in the half marathon. Since Half Marathon is less prestigious of a race, they are not sponsored by brands and are not forced to wear a specific shoe. Therefore, they just choose the shoe that is best. They are also younger so they are less acclimated to using the shoes that was deemed "best" in the past. Since they are younger and newer, half marathon runners do not have that deep of a connection to a certain model of shoe.

1500m

So how had middle distance changed over the years?

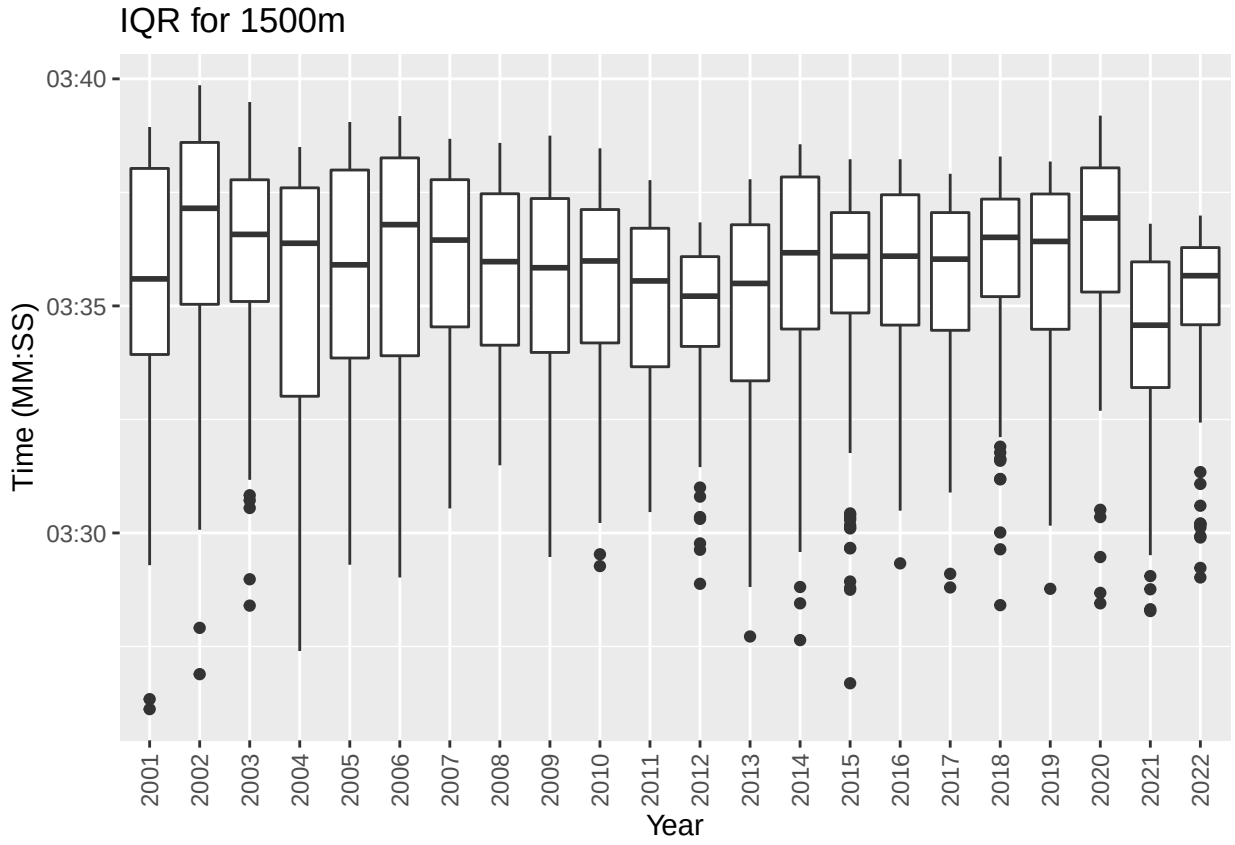
The 1500 is an interesting race since runners have different approaches towards it. Some treat it as a longer distance race while others treat it as shorter distance and prefer to wear spikes.

IQR

Let's run an IQR graph because there are different philosophies behind this race. Racers want to run the first half like a sprint and slow down or racers want to pace themselves out throughout the race. This variance in strategies cause the mean to be variant, however IQR should be more accurate. The mean would be skewed more heavily.

```
results %>%
  filter(Distance == "1500m", Gender == "M") %>%
  mutate(Year = substr(Date, 6, 9), Time = chron(times. = paste("00:", Mark, sep = ""))) %>%
  select(Competitor, Time, Year) %>%
  ggplot(aes(x=Year, y=Time)) + geom_boxplot() +
  scale_y_chron(format="%M:%S") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  ggtitle("IQR for 1500m") +
  ylab("Time (MM:SS)")
```

```
## Warning in convert.times(times., fmt): NAs introduced by coercion

## Warning in convert.times(times., fmt): 36 time-of-day entries out of range set
## to NA

## Warning: Removed 36 rows containing non-finite values (stat_boxplot).
```



Range of top values significantly increase after 2012, likely due to the release of the Nike Victory Elite shoe, demonstrating that carbon plated shoes could lead to a significant advantage for some runners. For a while, many continued to favor these shoes over even newer models through the 2016 release until 2021.

2021 can be seen to have a huge dip in IQR values, likely corresponding with the release of the revolutionary Nike Air Zoom Victory shoes

Fall of Q1 values reflect the later release of the Dragonfly spike and its inclusion in the meta for 1500 spikes. The shoe's slower rebound time and less aggressive design would justifiably lead to slower extremes

These shoes represented different approaches to the race, specifically those racers that decided to wear spikes.

800m

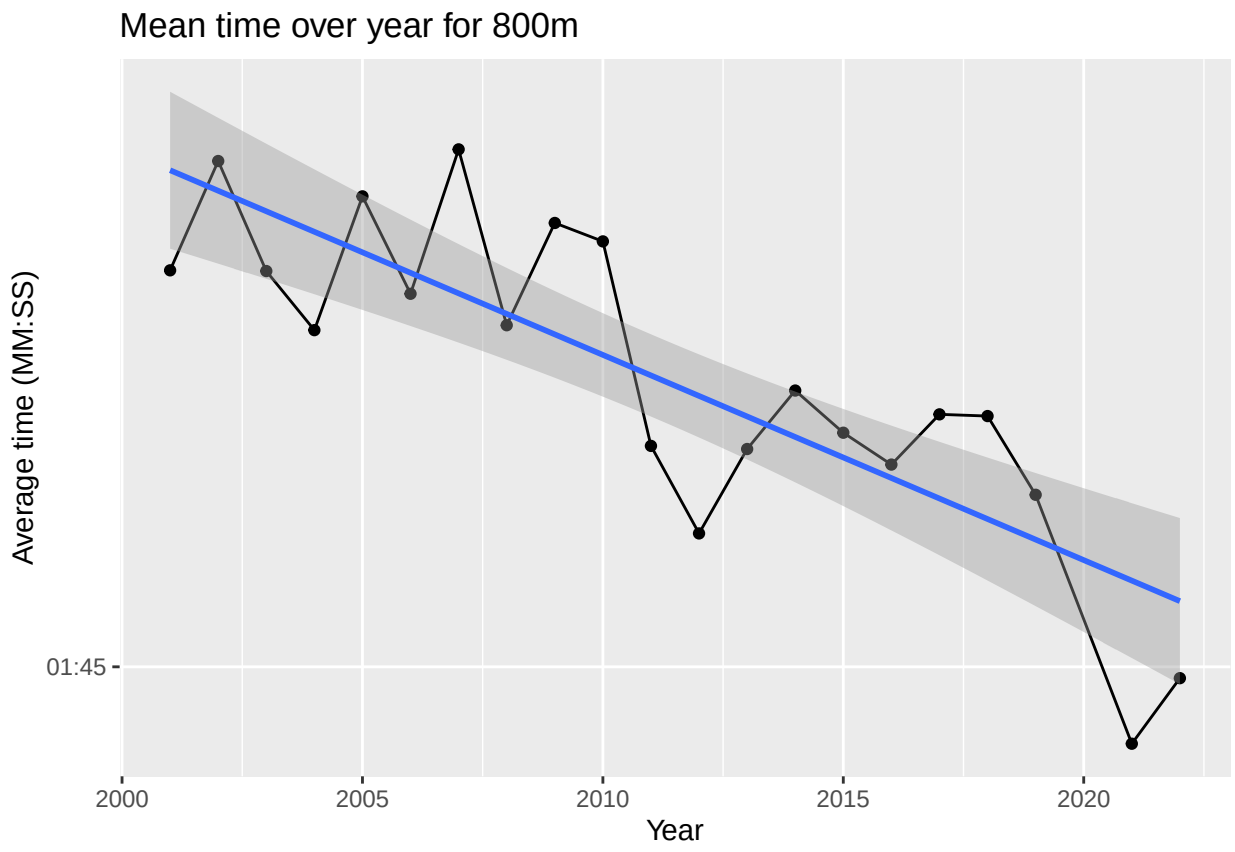
What does the times look for more middle distance races? Let's look at the 800m.

800m Mean time over year

There is less variation in strategy for the 800m so let's look at the mean times graph.

```
m800 %>%
  group_by(Year) %>%
  summarise(mean_time = mean(Time)) %>%
  arrange(mean_time) %>%
  ggplot(aes(x=Year, y=mean_time)) +
  geom_point() +
  geom_line() +
  geom_smooth(method = "lm") +
  scale_y_chron(format="%M:%S") +
  ggtitle("Mean time over year for 800m") +
  ylab("Average time (MM:SS)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



An analysis of the mean 800m times over the last 22 years reveals an apparent decrease in marks in 2012 with the release of the Nike Victory Elite. David Rudisha's time in the 2012 London Olympics of 1:40.91 has thought to be unbreakable and has led to people searching for new ways to push the pace, and opting for brands that had improved traction patterns, which did not work. This, however, ultimately proved ineffective in the following years

2016 saw the return of the carbon plated shoe in the Nike Victory Elite 2 (but plate was only 3/4 length) along with the newer 3d traction, leading to a slight decrease in lap times, but not nearly as much as 2012

2021 saw the introduction of the Air Zoom Victory, which features the maximum specifications allowed by World Athletics regulation (Zoom Air unit for optimal response time, Full length carbon internal plate, Higher Stack, Pliant External Plate), R&D found that for mid distance, and 3d traction. This year's new 3d traction, however, proved more of a disadvantage due to additional resistance and a lack of benefit of minute acceleration differences, so it was scrapped entirely in favor of new low resistance traction grooves which became commonly accepted as the best 800 spike in recent times, leading to times improving drastically

T test for 2010 and 2012, 2019 and 2021

What two year jump in technology was more significant?

2010-2012

```
m8002010 <- cleaned_results %>%
  filter(Distance=="800m", Gender=="M", Year==2010)

m8002012 <- cleaned_results %>%
  filter(Distance=="800m", Gender=="M", Year==2012)

tm800_2 <- m8002010 %>%
  rbind(m8002012)
t.test(Time ~ Year, data = tm800_2)
```

```
##
##  Welch Two Sample t-test
##
## data:  Time by Year
## t = 2.996529, df = 197.81, p-value = 0.0030787037037037
## alternative hypothesis: true difference in means between group 2010 and group 2012 is not equal to 0
## 95 percent confidence interval:
##  00:00:00 00:00:01
## sample estimates:
## mean in group 2010 mean in group 2012
##           00:01:46           00:01:45
```

2019-2021

```
m8002019 <- cleaned_results %>%
  filter(Distance=="800m", Gender=="M", Year==2019)

m8002021 <- cleaned_results %>%
  filter(Distance=="800m", Gender=="M", Year==2021)

tm800_1 <- m8002019 %>%
  rbind(m8002021)
t.test(Time ~ Year, data = tm800_1)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  Time by Year
## t = 3.300838, df = 183.86, p-value = 0.00115740740740741
## alternative hypothesis: true difference in means between group 2019 and group 2021 is not equal to 0
## 95 percent confidence interval:
##  00:00:00 00:00:01
## sample estimates:
## mean in group 2019 mean in group 2021
##           00:01:45           00:01:45
```

Both are statistically significant, due to having a p value below 0.05 The difference from 2010-2012 had a p value of 0.00307 meaning that probability that these results are due to random chance is low. Assuming there is no change in the general downward trend due to advancements in 800m-specific technology in 2010-2012, the probability of the graph ever taking on a value such as this randomly is astronomically low (as denoted by the very large t-test). The difference from 2019-2021 had a p value of 0.00115 meaning that probability that these results are due to random chance is low. Assuming there is no change in the general downward trend due to advancements in 800m-specific technology in 2019-2021, the probability of the graph ever taking on a value such as this randomly is astronomically low (as denoted by the very large t-test). However, the p value for 2019-2021 is lower meaning that this change was more significant.

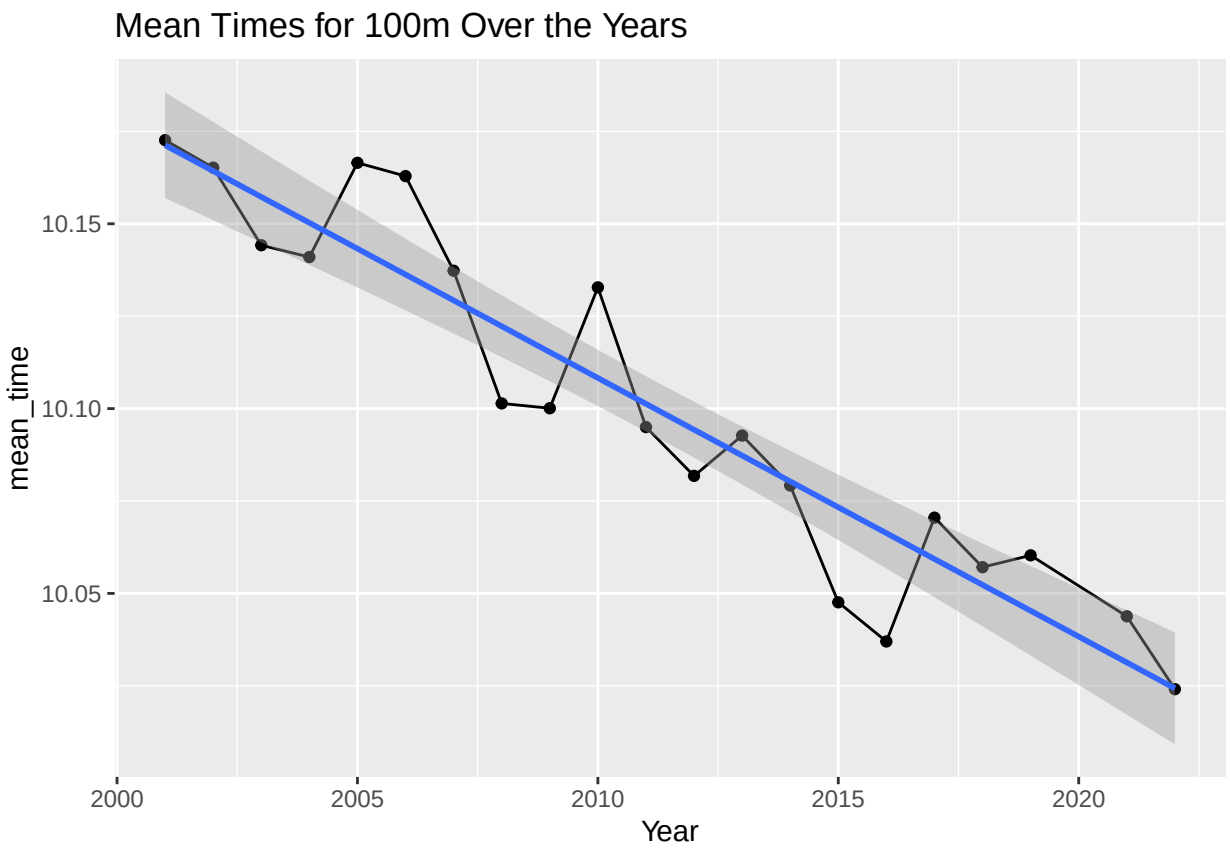
100m

100m Mean time over year

What do average 100m times look like over the last 21 years and how can we relate observations in trend-breaking points to advancements in sprint track and field technology?

```
results %>%
  filter(Distance == "100m", Gender == "M") %>%
  mutate(Year = strtoi(substr(Date,6,9))) %>%
  filter(Year !=2020) %>%
  group_by(Year) %>%
  summarise(mean_time = mean(as.numeric(Mark))) %>%
  arrange(mean_time) %>%
  ggplot(aes(x=Year, y=mean_time)) +
  geom_point() +
  geom_line() +
  geom_smooth(method = "lm") +
  ggtitle("Mean Times for 100m Over the Years")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
  ylab("time")
```

```
## $y
## [1] "time"
##
## attr(,"class")
## [1] "labels"
```

Significant improvements in race times in 2015-2016 can be attributed to developments in 3D traction technology for the Nike Superfly sprint shoes

Nike Maxfly (basically just an Air Zoom Victory but with more developmental focus on leg/ankle support over efficiency) released in 2021-2022, which led to a similar improvement in times

Not all 100m sprinters saw these technological advancements as beneficial; Usain Bolt has said that the new shoes are unfair and are bad for the sport, but as seen in the graph, there were spikes in improvement as people were experimenting with different shoe types.

Later versions of the Maxfly abandoned the 3d traction tech to remain under the stack limitation (height of "foam" looking part of the shoe), but traction has been proven to be extremely important for 100m

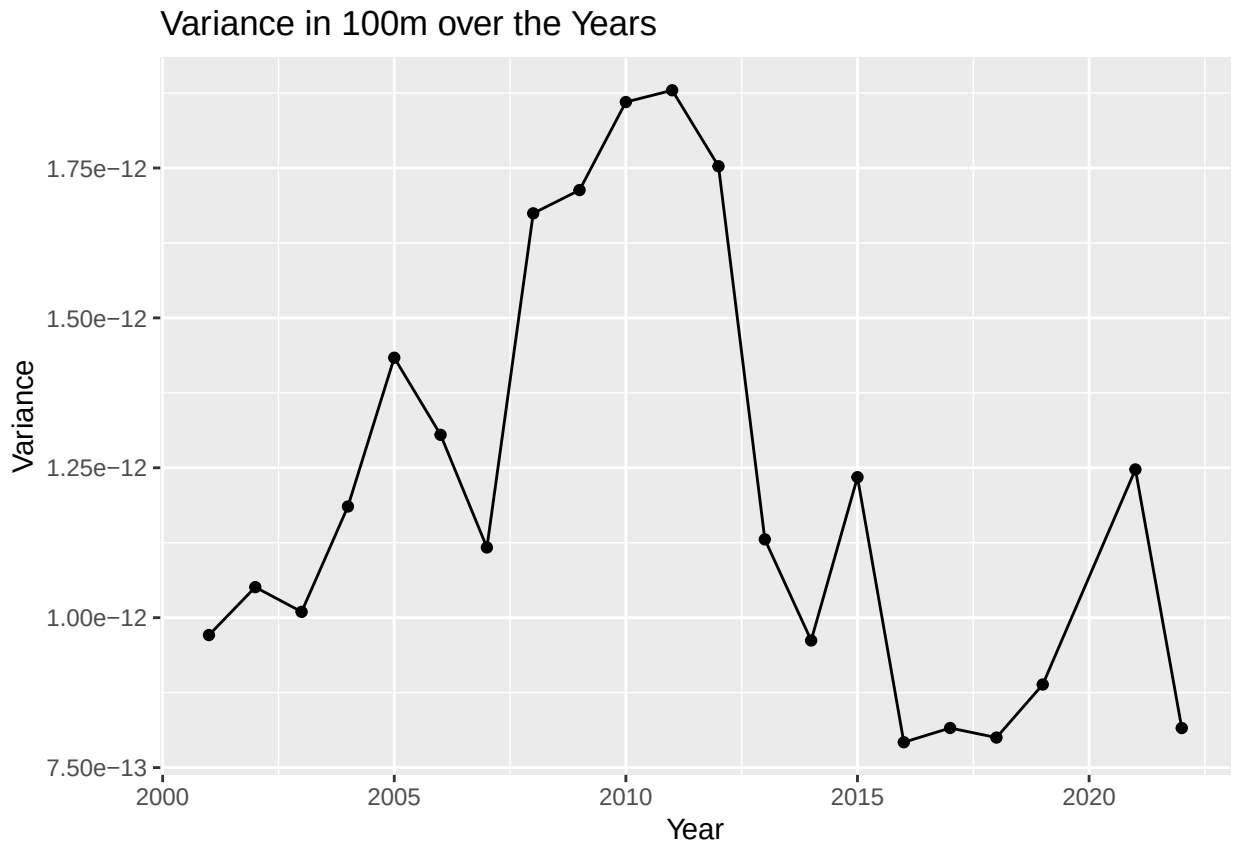
Both technological improvements led to undeniable improvements in times and if combined would have even more significant effects

Variance for 100m

Next, we turned towards analyses of variance in running times over the last 21 years. In the running world context, what causes contributed to spikes in variance?

Visualizing variance for sprint times of 100m races

```
cleaned_results %>%
  filter(Gender == "M", Distance == "100m") %>%
  group_by(Year) %>%
  summarize(variance = var(Time)) %>%
  ggplot(aes(x=Year, y=variance)) +
  geom_point() +
  geom_line() +
  ylab("Variance") +
  ggtitle("Variance in 100m over the Years")
```



It is seen that variance spikes in years that new shoes are released or not released entirely to the public. In 2012 it was the Nike Victory Elite spike that was released and the variance in the times prior reflect that. Then the spike in 2015 can be described by the Nike Superfly. It spikes down again as people started to adopt the shoes. Then again, there is spike in 2021 due to the Nike Maxfly when certain people decided to switch spikes while others did not.

Comparing 100m results to 200m results

Are improvements in sprint race times very similar? If not, why?

Note about 100m v 200m

The 100m and 200m races are vastly different as in the 100m, athletes often still have energy to run at their top speed when they finish the race. Therefore, the competition is about reaching the top speed faster and is all about acceleration, which is why better traction patterns are better as they let racers push into the ground with more energy. In contrast, the 200m is long enough where racers can reach their top speed and finish use all of the energy to maintain this top speed, so the race is about maintaining their top speed for the longest distance. This is why efficiency for a shoe matters so much more.

The following is a comparison of the improvement charts of 100m and 200m sprint race times. The purpose of these graphs is to create a side-by-side visual aid so that we may explore trends in 100m and 200m results as the years go on. If technology were to have little-to-no effect on the result trends, we would expect to see identical spikes and drops in race results over time as regulations change. However, if one graph features a unique/more exaggerated spike or drop not featured in the other, which as well coincides directly with major technological advancements in running shoes at the time, we can conclude that sudden changes in short-distance track and field running marks can be attributed to advancements in track and field technology.

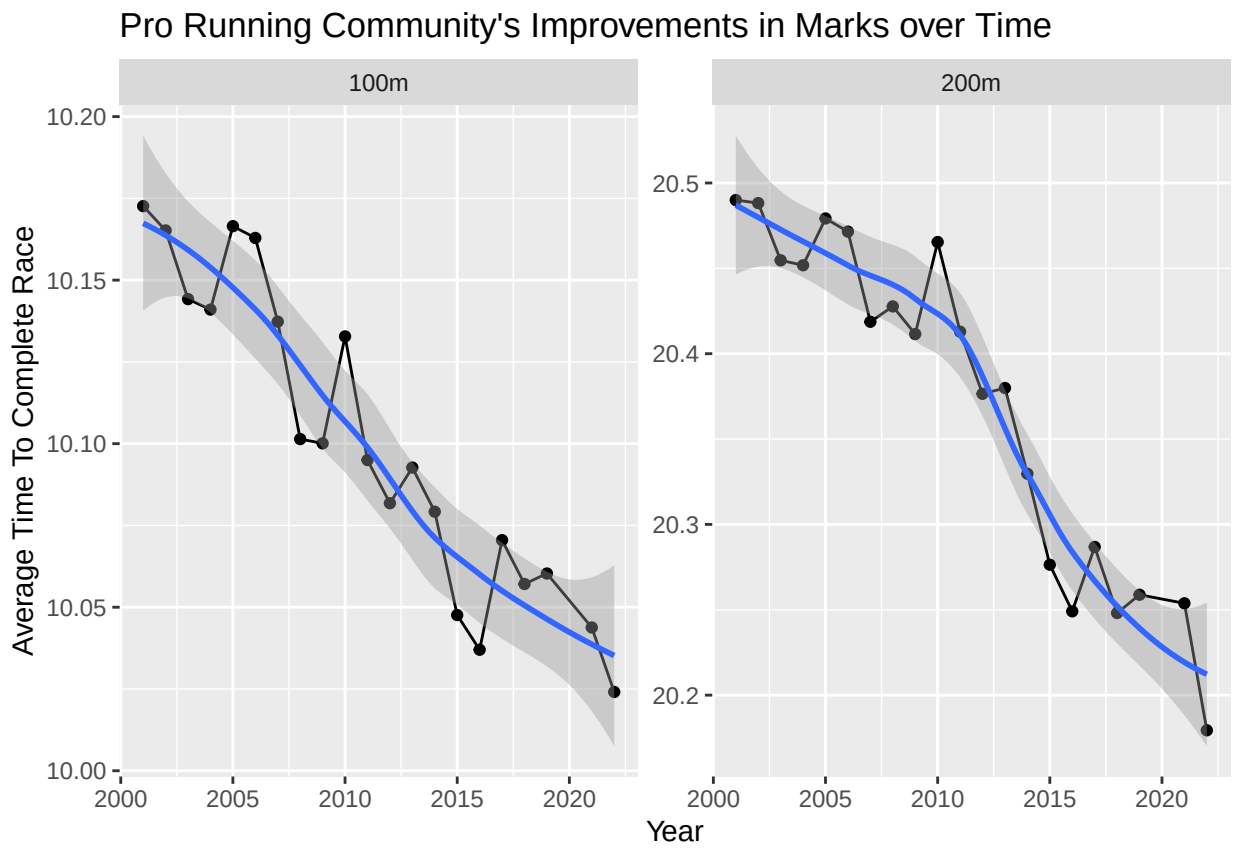
The below graph offers a side-by-side comparison of the two Mean Marks over Time graphs, this time with a LOESS model so that distinct graphic features like sudden spikes/drops can be observed more clearly

Improvements in 100m and 200m over time side by side

```
toGraph100 <- results %>%
  filter(Distance == "100m", Gender == "M") %>%
  mutate(Year = strtoi(substr(Date,6,9))) %>%
  filter(Year !=2020) %>%
  group_by(Year) %>%
  summarise(mean_time = mean(as.numeric(Mark))) %>%
  arrange(mean_time) %>%
    mutate(Dist = "100m")

toGraph200 <- results %>%
  filter(Distance == "200m", Gender == "M") %>%
  mutate(Year = strtoi(substr(Date,6,9))) %>%
  filter(Year !=2020) %>%
  group_by(Year) %>%
  summarise(mean_time = mean(as.numeric(Mark))) %>%
  arrange(mean_time) %>%
    mutate(Dist = "200m")

toGraph <- rbind(toGraph100, toGraph200)
```

```
ggplot(toGraph, aes(x = Year, y = mean_time)) +
  geom_point() +
  geom_line() +
  geom_smooth() +
  facet_wrap(~Dist, scale = "free") +
  labs(x = "Year", y = "Average Time To Complete Race", title = "Pro Running Community's Improvements i
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



In the graphic above, it is evident that there are pattern-breaking, drastic spikes downwards in marks in 2016 in both graphs. This is attributed to changes in technology regulations, which allowed for the use of new 3D traction in running spikes. More interestingly, however, is the similarly pattern-breaking decrease in times in 2021 for the 200m distance. We see that in the 200 event, times decreased in a much more exaggerated manner in 2021, which is due to the release of a new type of sprint spikes. These shoes were designed for 200m distances and combined short-distance sprint elements with the carbon fiber plates and zoom air traction of 800m shoes for the purpose of allowing runners to hold their maximum speed (cruising speed) for longer. These shoes were allowed to be used in 100m races, but they did not fit the event whatsoever as 100m shoes require much more grip to ensure athletes reach their top speed (in 200m distances, the concern is not whether atheltes reach their top speed or not, but for how long they can maintain it, hence the well-fittedness of this shoe to the 200m race distance).

Percent Change in 100m and 200m Times

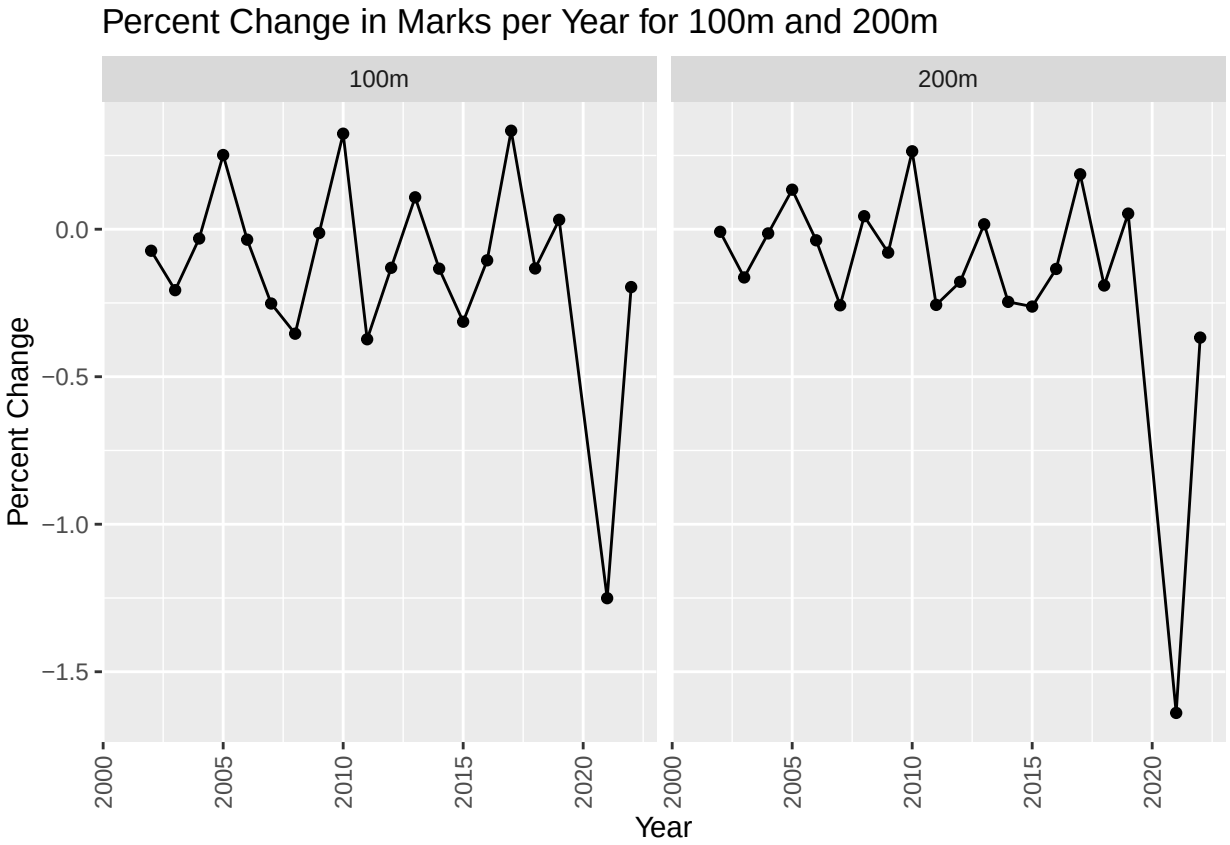
```
res100 <- results %>% filter(Distance == "100m") %>%
  filter(Gender == "M") %>%
    mutate(asNumMark = as.numeric(Mark)) %>%
    mutate(Year = strtoi(substr(Date, 6, 9)))

X100_pct <- res100 %>%
  filter(Gender == "M", Distance == "100m") %>%
  group_by(Year) %>%
```

```
  summarise(mean_time = mean(asNumMark)) %>%
  mutate(pct_change = ((mean_time) - lag(mean_time))/lag(mean_time)*100) %>%
  arrange(mean_time) %>%
  mutate(Event = "100m")

res200 <- results %>% filter(Distance == "200m") %>%
  filter(Gender == "M") %>%
    mutate(asNumMark = as.numeric(Mark))%>%
    mutate(Year = strtoi(substr(Date, 6, 9)))

X200_pct <- res200 %>%
  filter(Gender == "M", Distance == "200m") %>%
  group_by(Year) %>%
  summarise(mean_time = mean(asNumMark)) %>%
  mutate(pct_change = ((mean_time) - lag(mean_time))/lag(mean_time)*100) %>%
  arrange(mean_time) %>%
  mutate(Event = "200m")
```

```
X100_pct %>%
  rbind(X200_pct) %>%
  filter(Year != 2020) %>%
  ggplot(aes(x=Year, y=pct_change)) +
  geom_point() +
  geom_line() +
  facet_wrap(~Event) +
  ylab("Percent Change") +
  ggtitle("Percent Change in Marks per Year for 100m and 200m") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```



The visuals above simply reiterate the previous sentiments made in a more emphasized manner. It can be seen much clearer the relative difference in sprint marks in 2021. Moreover, it is made much clearer that the new sprint spikes released in 2021 impacted the 200m races much more than the 100m races as the percent change in 2021 for 200m is MUCH lower than the percent change in 2021 for 100m

Below we conducted by hand a one-sample t-test (we had to do it by hand and not using the t.test function because we needed to use results (instead of cleaned_results) since the sprint times are formatted in the more usable SS.MS (seconds, milliseconds) time format):

T test for 200m 2021

```
res200SUM <- res200 %>% mutate(Year = substr(Date, 6, 9)) %>%
  group_by(Year) %>%
    summarize(avgTime = mean(asNumMark))
```

```
xbar <- res200SUM$avgTime[21]
mu <- mean(res200$asNumMark, na.rm = TRUE)
n <- nrow(res200)
sdev <- sd(res200$asNumMark, na.rm = TRUE)
tVal <- (xbar - mu)/(sdev/sqrt(n))

tVal
```

```
## [1] -25.95398
```

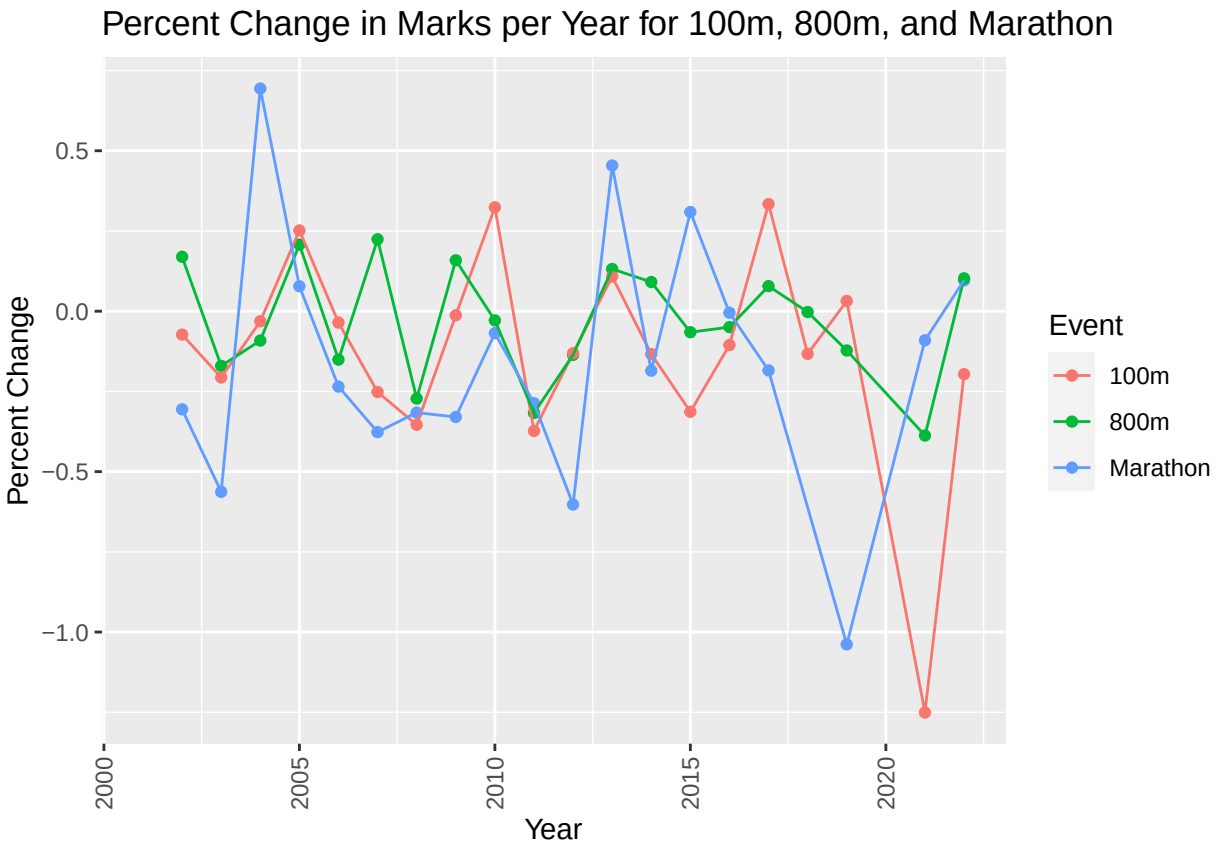
From this, we can conclude the following about the average mark for the year 2021 in the 200m distance graph: assuming there is no change in the general downward trend due to advancements in 200m-specific technology, the probability of the graph ever taking on a value such as this randomly is astronomically low (as denoted by the very large t-test).

Percent Change of Events at Each Distance Bracket (100m, 800m, Marathon)

```
resMarathon <- cleaned_results %>% filter(Distance == "Marathon") %>%
  filter(Gender == "M") %>%
    mutate(asNumMark = as.numeric(Time))

Marathon_pct <- resMarathon %>%
  group_by(Year) %>%
  summarise(mean_time = mean(asNumMark)) %>%
  mutate(pct_change = ((mean_time) - lag(mean_time))/lag(mean_time)*100) %>%
  arrange(mean_time) %>%
  mutate(Event = "Marathon")

res800m <- cleaned_results %>% filter(Distance == "800m") %>%
  filter(Gender == "M") %>%
    mutate(asNumMark = as.numeric(Time))

X800m_pct <- res800m %>%
  group_by(Year) %>%
  summarise(mean_time = mean(asNumMark)) %>%
  mutate(pct_change = ((mean_time) - lag(mean_time))/lag(mean_time)*100) %>%
  arrange(mean_time) %>%
  mutate(Event = "800m")

res100m <- results %>% filter(Distance == "100m") %>%
  filter(Gender == "M") %>%
    mutate(asNumMark = as.numeric(Mark)) %>%
    mutate(Year = strtoi(substr(Date, 6, 9)))

X100m_pct <- res100m %>%
  group_by(Year) %>%
  summarise(mean_time = mean(asNumMark)) %>%
  mutate(pct_change = ((mean_time) - lag(mean_time))/lag(mean_time)*100) %>%
  arrange(mean_time) %>%
  mutate(Event = "100m")

Marathon_pct %>%
  rbind(X800m_pct, X100m_pct) %>%
  filter(Year != 2020) %>%
  ggplot(aes(x=Year, y=pct_change, color=Event)) +
  geom_point() +
  geom_line() +
  ylab("Percent Change") +
  ggtitle("Percent Change in Marks per Year for 100m, 800m, and Marathon") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```

```
## Warning: Removed 3 row(s) containing missing values (geom_path).
```



We chose to overlay the percent change for only the marathon, 800m, and 100m as they encapsulate all different distances for races. If we were to overlay all races, the graph would become crowded and difficult to read. The 100m represents sprints, 800m represents middle distance, Marathon represents long distance and road. It shows that the 800 is more of a stable race and has less variation between years. The marathon has spikes and dips due to the longer training cycles of the racers.

Conclusion

Based on our analysis, years with significant drops in run times can be attributed to significant technological improvements. For example, in 2012 (middle distance), 2016 (sprints), 2019 (marathon), and 2021 (sprints and middle distance), our graphs showed that the significant drops in times happened in years where revolutionary technology in the form of shoes were released. In 2012 it was the Nike Zoom Victory Elite, then the in 2016 Nike Zoom Superfly, the Nike Vaporfly Next% in 2019, and the Nike Zoom Air Dragonfly in 2021. Therefore, we suspect there is validity in the claims of many professional athletes and companies that new shoe technology provides significant advantages.