**The Tour de France – a short primer**

The [Tour de France](https://en.wikipedia.org/wiki/Tour_de_France) (‘Le  
Tour’) is the world’s biggest and most prestigious cycling event with a  
long history spanning back as far as 1903. Each annual ‘edition’ of the  
race is composed of around 21 stages that traverse the French nation,  
each stage is a standalone race by itself. The racing is complex, with  
each team of 9 riders competing for any combination of individual stage  
wins, sprint points, mountain climbing, aggressive riding and team  
ability. The most coveted prize of all is the ‘Generale Classification’  
(GC) which is awarded to the rider with the lowest aggregate time at the  
end of the race. Each day, the rider with the lowest aggregate time  
following the previous stage wears the ‘Maillot Jaune’ (yellow jersey)  
indicating that they are the current race leader.

**tdf an R package for Tour de France data**

The tdf package contains  
information about the overall winning rider for each edition of the  
race, the winner’s biographical information and the results for each  
stage in each edition. To install the package, use

library(remotes)

The package is just a container for the dataframe editions:

library(tdf)

library(tidyverse)

# visualise contents tdf::editions

glimpse(editions)

## Observations: 106

## Variables: 20

## $ edition 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, …

## $ start\_date 1903-07-01, 1904-07-02, 1905-07-09, 1906-07-04, 1907-0…

## $ winner\_name "Maurice Garin", "Henri Cornet", "Louis Trousselier", "…

## $ winner\_team "La Française", "Conte", "Peugeot–Wolber", "Peugeot–Wol…

## $ distance 2428, 2428, 2994, 4637, 4488, 4497, 4498, 4734, 5343, 5…

## $ time\_overall 94.55389, 96.09861, NA, NA, NA, NA, NA, NA, NA, NA, 197…

## $ time\_margin 2.98916667, 2.27055556, NA, NA, NA, NA, NA, NA, NA, NA,…

## $ stage\_wins 3, 1, 5, 5, 2, 5, 6, 4, 2, 3, 1, 1, 1, 4, 2, 0, 3, 4, 4…

## $ stages\_led 6, 3, 10, 12, 5, 13, 13, 3, 13, 13, 8, 15, 2, 14, 14, 3…

## $ height 1.62, NA, NA, NA, NA, NA, 1.78, NA, NA, NA, NA, NA, NA,…

## $ weight 60, NA, NA, NA, NA, NA, 88, NA, NA, NA, NA, NA, NA, NA,…

## $ age 32, 19, 24, 27, 24, 25, 22, 22, 26, 23, 23, 24, 33, 30,…

## $ born 1871-03-03, 1884-08-04, 1881-06-29, 1879-06-05, 1882-1…

## $ died 1957-02-19, 1941-03-18, 1939-04-24, 1907-01-25, 1917-1…

## $ full\_name NA, NA, NA, NA, "Lucien Georges Mazan", "Lucien Georges…

## $ nickname "The Little Chimney-sweep", "Le rigolo (The joker)", "L…

## $ birth\_town "Arvier", "Desvres", "Paris", "Moret-sur-Loing", "Pless…

## $ birth\_country "Italy", "France", "France", "France", "France", "Franc…

## $ nationality " France", " France", " France", " France", " France", …

## $ stage\_results [[, ,

editions is a tibble whose rows each correspond to a single edition of  
the Tour de France. The columns contain information about the race  
itself and the overall winner, including:

* distance is the aggregate distance in kilometres covered by the  
  entire race.
* time\_overall is the time in hours taken by the winner to complete  
  the race
* time\_margin is the difference in finishing times between the race  
  winner and the first runner up.
* stage\_wins is the number of stages won by the eventual winner  
  during the edition (note that it is possible to win the GC without  
  winning any stages at all).
* stages\_led is the number of stages spent as the race leader  
  (wearing the yellow jersey) by the eventual winner.
* weight is the winner’s body weight in kilograms.
* height is the winner’s height in meters.
* stage\_results is a column containing a list of lists. Each element  
  contains a list of stage results for a particular edition of the  
  Tour de France.

**How has the race changed over time?**

Forget ultra-marathons and tough mudder, early editions of Le Tour were  
**really tough**. Riders were mostly self-supported, rode in woollen  
jerseys for hundreds of miles per day on steel-framed bicycles. The  
longest stage in Tour history was 482 kilometres (Stage 5, 1919) – the  
stage winner, Jean Alavoine, took almost 19 hours to complete the stage.

\*Robert Jacquinot taking a break from the 482km long Stage 5 during the  
1922 Tour de France. (Image source: Wikipedia.)\*

To get a sense for how the length of the race has varied since 1903, we  
can visualise the total distance in the editions data:

library(ggplot2)

editions %>%

ggplot(aes(x = start\_date, y = distance,

color = edition)) +

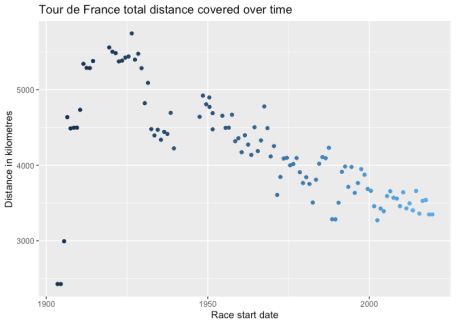
geom\_point() +

xlab('Race start date') +

ylab('Distance in kilometres') +

ggtitle('Tour de France total distance covered over time') +

theme(legend.position = "none")



It’s pretty clear that over time, the distances covered have decreased  
dramatically, and have roughly stabilised at about 3500 kilometres  
during the last 2 decades (still a huge distance). You can see that the  
longest ever Tour de France edition was in 1926, with a total distance  
covered of 5,745 kilometres!

On the face of it, it seems like the riders of today have it  
substantially easier compared to riders of the past. But how fast are  
today’s riders going?

library(ggrepel)

editions %>%

ggplot(aes(x = start\_date, y = distance / time\_overall,

color = edition)) +

geom\_point(na.rm = TRUE) +

geom\_label\_repel(data = editions %>% sample\_n(20),

aes(label = winner\_name), size = 2.3,

nudge\_y = -9, na.rm = TRUE,

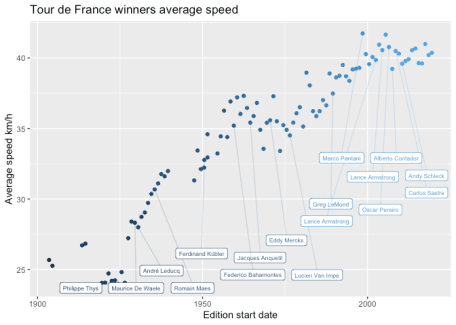
segment.alpha = 0.2) +

xlab('Edition start date') +

ylab('Average speed km/h') +

ggtitle('Tour de France winners average speed') +

theme(legend.position = "none")



They’re going pretty fast. It looks like while the race has been getting  
gradually shorter, the speeds have been getting much faster. The change  
also coincides with professionalisation of the sport, better equipment  
and smarter training so it’s hard to provide an exact account for the  
change in speed. It’s worth highlighting the top two fastest average  
speeds in Tour de France history:

# Top 5 average speeds of Tour de France winners

editions %>%

mutate(speed = distance / time\_overall) %>%

select(start\_date, winner\_name, speed) %>%

arrange(desc(speed)) %>%

print(n = 2)

## # A tibble: 106 x 3

## start\_date winner\_name speed

##

## 1 1998-07-11 Marco Pantani 41.7

## 2 2005-07-02 Lance Armstrong 41.7

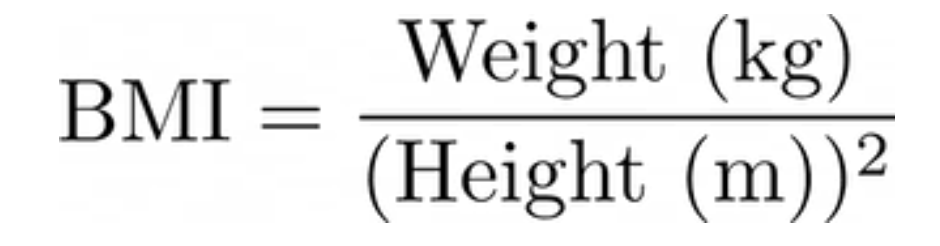
## # … with 104 more rows

The two fastest ever editions of the Tour de France were won by Marco  
Pantani (in 1998) and Lance Armstrong (in 2005), both of whom were later  
stripped of these (and other) wins for their use of banned  
performance-enhancing substances. The speed of doped riders in such Tour  
editions was so obviously faster than non-doped riders, that French  
media declared a culture of *“Cyclisme a deux vitesses”* (*“two-speed  
cycling”*). It is unknown how much riders still use banned substances  
for performance enhancement, but the average speeds of the Pantani /  
Armstrong years have not been reached in any edition since.

**Note:** the the data in the tdf package retains the winning times of  
banned, disqualified and otherwise sanctioned riders for the purposes of  
data analysis. The overall standings are as they would have appeared on  
the final day of the race – therefore please note that the officially  
recognised winner of a particular edition may not be the rider with the  
fastest time.

**How have the riders changed over time?**

France is a mountainous country, and a crucial ingredient for success in  
the Tour de France is a rider’s ability to climb hills quickly and  
efficiently. Hill climbing is a fight against gravity that pits a  
rider’s strength against their total weight (bike + equipment + body).  
The rider has two options to improve: get stronger and get leaner. Using  
the editions data we can explore the latter over time by using rider  
height and weight data to calculate body mass index (BMI), which is  
a (very rough) proxy for leanness.



library(ggrepel)

editions %>%

ggplot(aes(x = start\_date, y = weight / height^2,

color = edition)) +

geom\_point(na.rm = TRUE) +

geom\_label\_repel(data = editions %>% sample\_n(20),

aes(label = winner\_name), size = 1.8,

nudge\_y = 4, na.rm = TRUE,

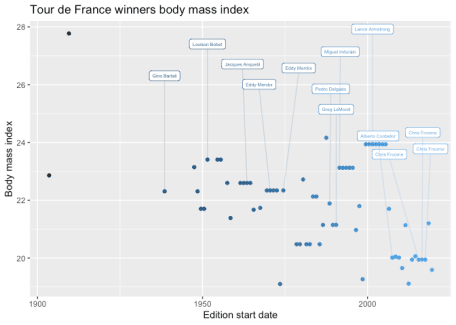
segment.alpha = 0.2) +

xlab('Edition start date') +

ylab('Body mass index') +

ggtitle('Tour de France winners body mass index') +

theme(legend.position = "none")



It’s pretty clear that over time, the trend has been towards winners  
having lower BMI, and likely being leaner overall. Apart from the  
obvious issues with BMI as a metric (body shapes are more complex than  
just height and weight) it’s interesting to consider why this trend has  
occurred. It’s tempting to conclude that more careful dieting and  
preparation in recent years has lead to riders having lower body fat  
percentages, which can enhance a rider’s power to weight ratio and  
overall performance. However, it could also be due to changes in the  
race: if race winning becomes more dependent on performance in the  
mountains (for example, because the number of mountain stages has  
increased overall) this could result in the lighter and leaner athletes  
tending to excel overall.

**Stage results**

The column stage\_results contains the breakdown of results by stage  
for each edition of the Tour de France. For example, the results of the  
final stage of the 2019 Tour de France can be printed using

editions$stage\_results$`2019`$`stage-21`

## # A tibble: 155 x 8

## rank time rider bib\_number age team points elapsed

##

## 1 1 3H 4M 8S Ewan Caleb 161 25 Lotto Soudal 100 3H 4M 8S

## 2 2 0S Groenewegen … 84 26 Team Jumbo-Vis… 70 3H 4M 8S

## 3 3 0S Bonifazio Ni… 172 25 Team Total Dir… 50 3H 4M 8S

## 4 4 0S Richeze Maxi… 27 36 Deceuninck - Q… 40 3H 4M 8S

## 5 5 0S Boasson Hage… 201 32 Team Dimension… 32 3H 4M 8S

## 6 6 0S Greipel André 215 37 Team Arkéa Sam… 26 3H 4M 8S

## 7 7 0S Trentin Matt… 107 29 Mitchelton-Sco… 22 3H 4M 8S

## 8 8 0S Stuyven Jasp… 138 27 Trek - Segafre… 18 3H 4M 8S

## 9 9 0S Arndt Nikias 142 27 Team Sunweb 14 3H 4M 8S

## 10 10 0S Sagan Peter 11 29 BORA - hansgro… 10 3H 4M 8S

## # … with 145 more rows

The important columns for the stage data are

* time the finishing time of the stage winner and time difference to  
  the winner.
* rider the rider name formatted as ‘Surname Forename’.
* age age of the rider at the start of the stage.
* elapsed the time taken to reach the finish line – this is stored  
  as a lubridate::period object for easier printing and  
  manipulation.

In the case above, Caleb Ewan won the finish line sprint of the final  
stage. Since the first 53 riders were part of a contiguous group of  
riders, they were granted the same finishing time as Ewan, but their  
finishing order corresponds to the order they passed the finish line.

\*Caleb Ewan wins the sprint on stage 21 of the 2019 Tour de France.  
(Image source: AFP/Getty Images)\*