Marathon Data

Sturdivant and Clark

2022-08-03

Getting Data

Code to get data from wikipedia

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

```
## Warning: package 'purrr' was built under R version 4.0.5
```

```
## Warning: package 'stringr' was built under R version 4.0.5
```

```
## Warning: package 'forcats' was built under R version 4.0.5
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(rvest)
```

```
## Warning: package 'rvest' was built under R version 4.0.5
```

```
##
## Attaching package: 'rvest'
```

```
## The following object is masked from 'package:readr':
##
##     guess_encoding
```

```r
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.0.5
```

```
##
## Attaching package: 'lubridate'
```

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

```r
theme_set(theme_bw())

url = "https://en.wikipedia.org/wiki/Marathon_world_record_progression"

marathon_html = read_html(url)

marathon_html %>%
  html_nodes(css = "table")
```

```
## {xml_nodeset (5)}
## [1] <table class="wikitable" style="font-size: 95%;"><tbody>\n<tr style="back ...
## [2] <table class="wikitable" style="font-size: 95%;"><tbody>\n<tr style="back ...
## [3] <table class="nowraplinks mw-collapsible autocollapse navbox-inner" style ...
## [4] <table class="nowraplinks navbox-subgroup" style="border-spacing:0"><tbod ...
## [5] <table class="nowraplinks hlist mw-collapsible autocollapse navbox-inner" ...
```

```r
record_table =
  marathon_html %>%
  html_nodes(css = "table") %>%
  nth(1) %>%
  html_table(fill = TRUE)

print(record_table)
```

```
## # A tibble: 50 x 7
##    Time      Name            Nationality    Date        Event~1 Source Notes
##    <chr>     <chr>           <chr>          <chr>       <chr>   <chr>  <chr>
##  1 2:55:18.4 Johnny Hayes    United States  July 24, 19~ London~ IAAF[~ "Tim~
##  2 2:52:45.4 Robert Fowler   United States  January 1, ~ Yonker~ IAAF[~ "Not~
##  3 2:46:52.8 James Clark     United States  February 12~ New Yo~ IAAF[~ "Not~
##  4 2:46:04.6 Albert Raines   United States  May 8, 1909  New Yo~ IAAF[~ "Not~
##  5 2:42:31.0 Henry Barrett   United Kingdom May 8, 1909~ Polyte~ IAAF[~ "Not~
##  6 2:40:34.2 Thure Johansson Sweden         August 31, ~ Stockh~ IAAF[~ "Not~
##  7 2:38:16.2 Harry Green     United Kingdom May 12, 1913 Polyte~ IAAF[~ "Not~
##  8 2:36:06.6 Alexis Ahlgren  Sweden         May 31, 1913 Polyte~ IAAF[~ "Rep~
##  9 2:38:00.8 Umberto Blasi   Italy          November 29~ Legnan~ ARRS[~ ""
## 10 2:32:35.8 Hannes Kolehmainen Finland     August 22, ~ Antwer~ IAAF,~ "The~
## # ... with 40 more rows, and abbreviated variable name 1: `Event/Place`
## # i Use `print(n = ...)` to see more rows
```

```
record_table_mod =
  record_table %>%
  mutate(Time_t = hms(Time))%>%
  mutate(Time_sec = period_to_seconds(Time_t))

diff(record_table_mod$Time_t)
```

```
##  [1]  27.0   7.4 -48.2  26.4   3.2 -18.0  -9.6  -5.8  35.0 -34.0  55.8 -43.6
## [13]  35.0  -5.0  -2.0  -3.0   3.2  -1.8  -5.6   4.6 -34.6  12.2  -0.8  -0.4
## [25]  12.2  15.0  12.0 -43.8 -11.2  36.4  -2.8  -4.8 -16.8  -6.4  -4.6  17.0
## [37] -13.0   7.0  38.0 -45.0  37.0  -4.0  17.0 -29.0  33.0 -21.0 -15.0  34.0
## [49] -18.0
## attr(,"class")
## [1] "Period"
## attr(,"class")attr(,"package")
## [1] "lubridate"
```

```
record_table_mod =
  record_table_mod %>%
  mutate(Date_ymd = mdy(Date))
```
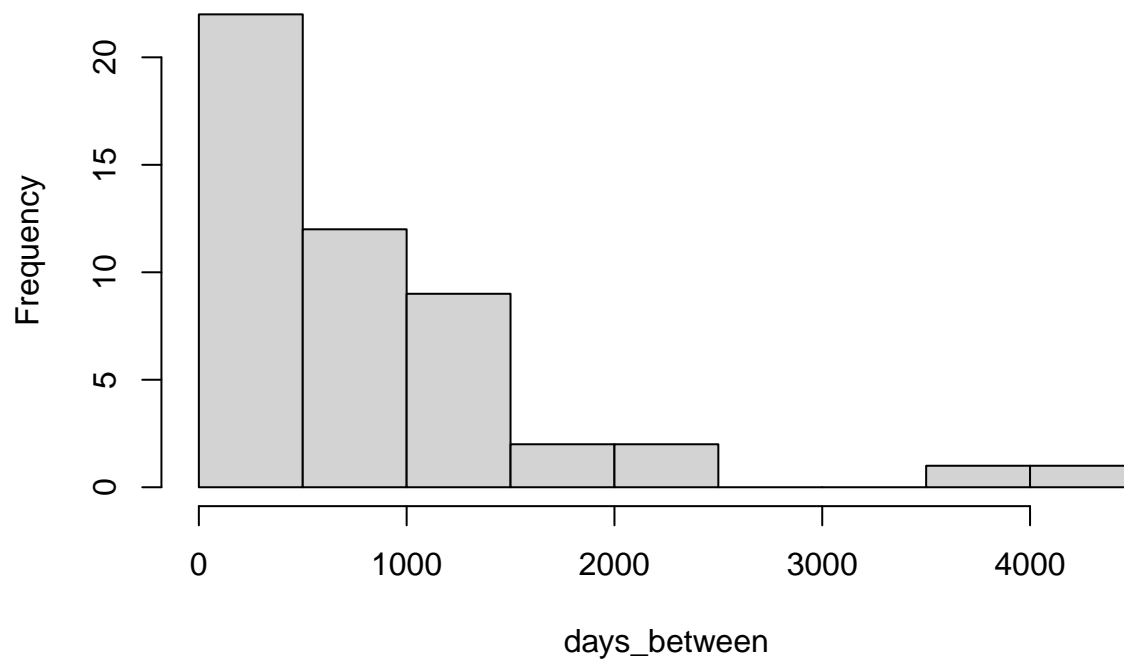
```
## Warning: 1 failed to parse.
```

```
record_table_mod$Date_ymd[5] = "1909-05-10"

days_between = as.numeric(diff(record_table_mod$Date_ymd))
```
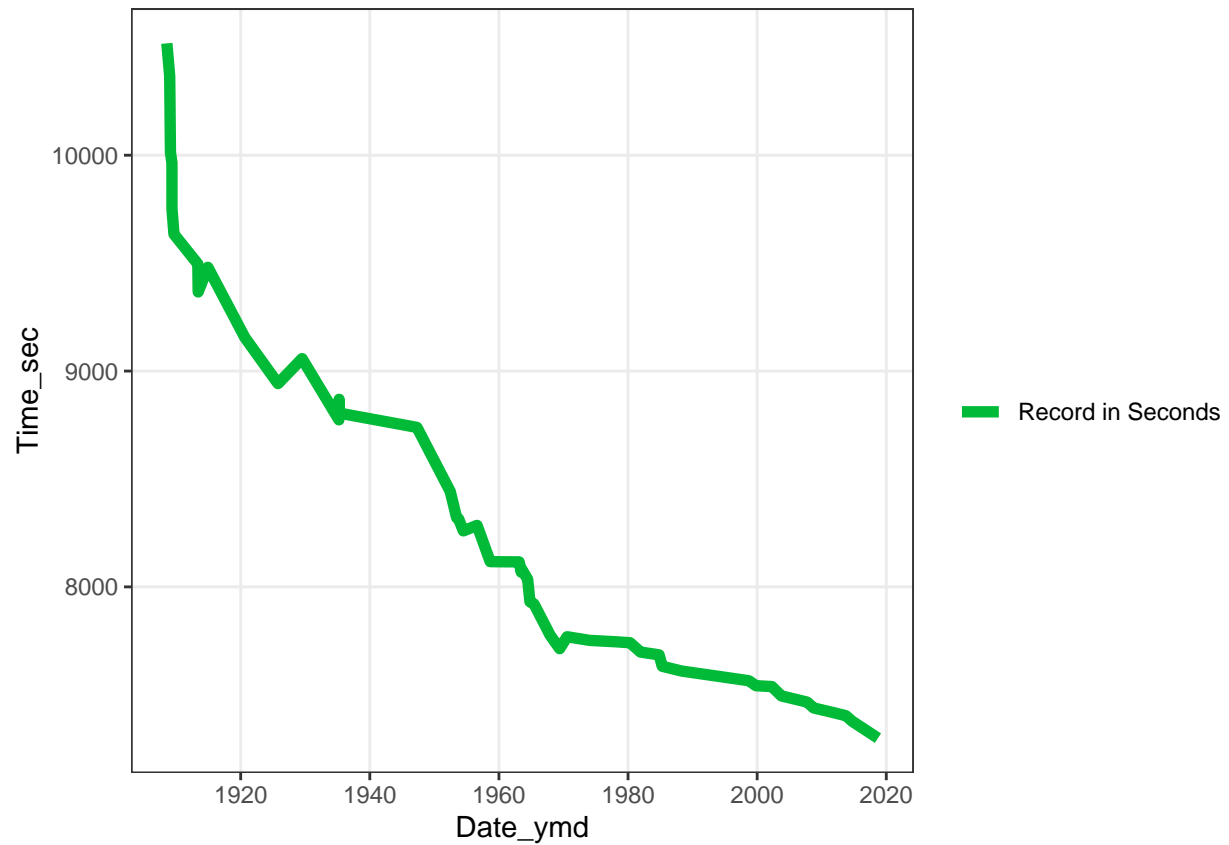
```
hist(days_between)
```

Histogram of days_between


```
record_table_mod %>% ggplot() +
  geom_line(aes(x=Date_ymd, y=Time_sec, color = "Record in Seconds"),size=2) +
  scale_color_manual(name="",
                     values = c("Record in Seconds"="#00ba38")) +
  theme(panel.grid.minor = element_blank())
```

```
record_table_mod$end = c(record_table_mod$Date_ymd[-1],"2022-08-01")


ggplot() +
  geom_segment(data=record_table_mod, aes(x=Date_ymd,xend=end,y=Time_sec,yend=Time_sec),size=2) +
  ylab("Time in Seconds")+
  xlab("Marathon Record")+
  geom_rug(data=record_table_mod,aes(x=Date_ymd,y=Time_sec,),sides="b")
```