JOHN KING'ATHIA KARUITHA

WHO'S THE FASTEST OF ALL? ANALYZING THE 100 METRES MEN'S SPRINT DATA, 1900-2021

Background

I use data from World Athletics on the best times posted by male athletes in the 100 metres sprint from 1900 to present [1]. Note that many athletes appear multiple times provided the times. For instance, Usain Bolt posted top 3 best times in different races.

[1] The data is available on this link `https://www.worldathletics.org/records/all-time-toplists/sprints/100-metres/outdoor/men/senior?regionType=world&timing=electronic&windReading=regular&page=21&bestResultsOnly=false&firstDay=1900-01-01&lastDay=2021-09-20`

```
read_chunk("code/sprints.R")
```

The raw data consists of the following variables;

- rank: Starting from the athlete that has posted the best time to date.
- mark: The time (in seconds) posted by the athlete.
- wind: The wind assist. Negative speeds indicate the athlete was running against the wind.
- competitor: The name of the athlete.
- dob: Date of birth of the athlete.
- nat: Nationality of the athlete.
- pos: position of the athlete in the given race.
- venue: Venue of the race.
- date: Date the race happened.
- results_score: The score of the athlete in the race by World Athletics.

The Data

As noted, the data is available in the World Athletics website. This data spans two hundred and twenty four (224) web pages at the time of this write up. Tellingly, it would take ages to copy-paste this data. Hence, the first exercise is to scrap the data. Please refer to my previous project on web scrapping available on `https://rpubs.com/Karuitha/web_scrapping_1`. The steps in scrapping the data are outlined below.

Web Scrapping

The scrapping process takes considerable time. For this reason, I have commented out the last two lines of code that do the scrapping.

NOTE: To repeat the scrapping process, just remove the # before the two lines of code.

```
pages <- 1:224


url <- "https://www.worldathletics.org/records/all-time-toplists/sprints/100-metres/outdoor/men/senior?

url_2 <- "&bestResultsOnly=false&firstDay=1900-01-01&lastDay=2021-09-20"

## Full url example
full_url <- paste0(url, "1", url_2)

## Scrapping function
scrapper <- function(x){

        Sys.sleep(4)

        read_html(paste0(url, x, url_2)) %>%

                html_nodes("table") %>%

                html_table()

}



# my_100_dash_data <- pages %>% map_dfr(~ scrapper(.x))

# write_csv(my_100_dash_data, "my_100_dash_data.csv")
```

Data Cleaning and Feature Engineering

The resultant dataset has 22,400 rows and 14 columns. The data cleaning process involves converting the date of `birth` (dob) and `date` to date/ time format. After this, I did feature engineering, adding the following variables.

- Age of athletes at the time of the race in days.
- Age of athletes at the time of the race in years. I divided the age in days by 365.25 to get years.

- Venue country code: The code of the country where the race happened.
- Venue country name: I used the `countrycode` package in R to convert the country codes into country names. Where missing, I used information available on `https://www.olympiandatabase.com/index.php?id=1670&L=1` to fill in the country names.

```
my_100_dash_data <- read_csv("data/my_100_dash_data.csv") %>%
        clean_names() %>%
        select(-x8) %>%
        mutate(dob = lubridate::dmy(dob),
               date = lubridate::dmy(date),
               age_days = (date - dob),
               age_years = as.numeric(age_days / 365.25),
               venue_country_code = str_extract_all(venue, "\\([A-Z]*\\)"),
               venue_country_code = str_remove_all(venue_country_code, "\\(|\\)")) %>%
        mutate(venue_country_name = countrycode(venue_country_code,
                                                origin = "ioc",
                                                destination = "country.name")) %>%
        mutate(venue_country_name = case_when(venue_country_code == "AHO" ~ "Netherlands Antilles",
                                              venue_country_code == "FRG" ~ "Germany",
                                              venue_country_code == "GDR" ~ "Germany",
                                              venue_country_code == "MAC" ~ "Macau",
                                              venue_country_code == "TCH" ~ "Czechia",
                                              venue_country_code == "TKS" ~ "Turks and Caicos Islands",
                                              venue_country_code == "URS" ~ "Russia",
```

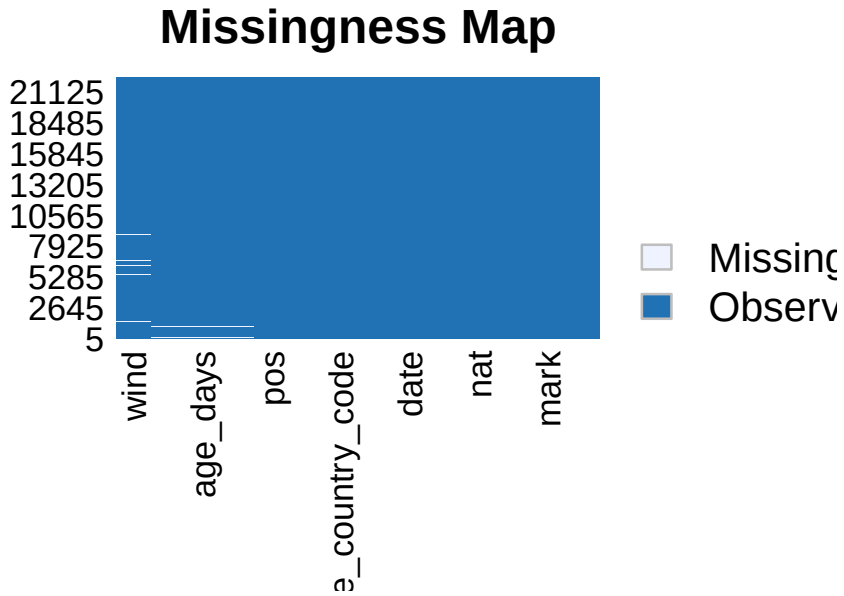
```
                                        TRUE ~ venue_country_name

                                        ))
```

Missing Data

Only 4 variables have missing data- wind, dob, age_days, age_years, and pos. However, the extent of missingness is not high as the proportion of missing data shows in the table below.

```
Amelia::missmap(my_100_dash_data)
```




```
sapply(my_100_dash_data, is.na) %>%

        colSums() %>%

        tibble(variables = names(my_100_dash_data), missing = .) %>%

        arrange(desc(missing)) %>%

        mutate(prop_percent = missing / nrow(my_100_dash_data) * 100) %>%

        head(10)
```

```
## # A tibble: 10 x 3
##    variables  missing prop_percent
##    <chr>        <dbl>        <dbl>
##  1 wind           431        1.92
```

```
##  2 dob            181        0.808
##  3 age_days       181        0.808
##  4 age_years      181        0.808
##  5 pos             50        0.223
##  6 rank             0        0
##  7 mark             0        0
##  8 competitor       0        0
##  9 nat              0        0
## 10 venue            0        0
```

Exploratory Data Analysis

I examine the country with the most athletes that have posted the best times in the 100 metres dash. The United States leads as the table below shows. However, as noted ealier, the athletes are repeated given that one athlete may have posted multiple times. In table () I remove the duplicates to get the country with the most athletes which again is the United States.

Most Successful Countries

```
my_100_dash_data %>%

        count(nat, sort = TRUE)
```

```
## # A tibble: 127 x 2
##    nat       n
##    <chr> <int>
##  1 USA    7086
##  2 JAM    2426
##  3 GBR    1648
##  4 NGR    1076
##  5 CAN     887
##  6 JPN     736
##  7 TTO     729
##  8 FRA     600
##  9 RSA     555
## 10 BRA     499
## # ... with 117 more rows
```

```
my_100_dash_data %>%

        select(competitor, nat) %>%

        filter(!duplicated(.)) %>%
```

```
        count(nat) %>%

        arrange(desc(n))
```

```
## # A tibble: 127 x 2
##    nat       n
##    <chr> <int>
##  1 USA     691
##  2 JAM     132
##  3 GBR      90
##  4 JPN      90
##  5 NGR      71
##  6 CAN      55
##  7 RSA      52
##  8 CHN      47
##  9 FRA      44
## 10 GER      42
## # ... with 117 more rows
```

World Record Holders

The table below shows the 10 athletes who have posted the best times in the 100 metres dash. Note that Usain Bolt appears in this list 4 times.

```
my_100_dash_data %>%

        select(competitor, mark) %>%

        arrange(mark) %>%

        head(10)
```

```
## # A tibble: 10 x 2
##    competitor     mark
##    <chr>         <dbl>
##  1 Usain BOLT     9.58
##  2 Usain BOLT     9.63
##  3 Usain BOLT     9.69
##  4 Tyson GAY      9.69
##  5 Yohan BLAKE    9.69
##  6 Tyson GAY      9.71
##  7 Usain BOLT     9.72
##  8 Asafa POWELL   9.72
##  9 Asafa POWELL   9.74
```

```
## 10 Justin GATLIN  9.74
```

Top 10 100 Metres Dash Athletes

It is common knowledge that Usain Bolt is the record holder. But who are the other top contenders. Usain Bolt holds the three best times in the dash. To deal with this duplication, I remove duplicates so that we have one entry per athlete. With that, the top 10 athletes are shown below.

```
my_100_dash_data %>%

        select(competitor, mark) %>%

        group_by(competitor) %>%

        arrange(mark) %>%

        slice(1) %>%

        ungroup() %>%

        arrange(mark) %>%

        head(10)
```

```
## # A tibble: 10 x 2
##    competitor          mark
##    <chr>              <dbl>
##  1 Usain BOLT          9.58
##  2 Tyson GAY           9.69
##  3 Yohan BLAKE         9.69
##  4 Asafa POWELL        9.72
##  5 Justin GATLIN       9.74
##  6 Christian COLEMAN   9.76
##  7 Trayvon BROMELL     9.76
##  8 Ferdinand OMANYALA  9.77
##  9 Nesta CARTER        9.78
## 10 Maurice GREENE      9.79
```

Athletes With the Most appearances in the Fastest Athletes List

The issue here is to examine the athlete that has appeared in the list of elite athletes the most times. Here, Michael Rodgers from the USA leads the way having run 267 races with some of the best times in the World.

```
my_100_dash_data %>%

        count(competitor, sort = TRUE) %>%

        head(10)
```

```
## # A tibble: 10 x 2
##    competitor           n
##    <chr>            <int>
##  1 Michael RODGERS    267
##  2 Kim COLLINS        224
##  3 Asafa POWELL       196
##  4 Dennis MITCHELL    192
##  5 Michael FRATER     182
##  6 Frank FREDERICKS   173
##  7 Justin GATLIN      162
##  8 Francis OBIKWELU   161
##  9 Linford CHRISTIE   161
## 10 Bruny SURIN        155
```