Assignment 5

### Load the dataset

#load the dataset  
library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.1 ✔ readr 2.1.4  
✔ forcats 1.0.0 ✔ stringr 1.5.0  
✔ ggplot2 3.4.1 ✔ tibble 3.2.1  
✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
✔ purrr 1.0.1   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(here)

here() starts at C:/Users/calis/OneDrive/Desktop/BUAN4210/R

library(cowplot)

Attaching package: 'cowplot'  
  
The following object is masked from 'package:lubridate':  
  
 stamp

library(scales)

Attaching package: 'scales'  
  
The following object is masked from 'package:purrr':  
  
 discard  
  
The following object is masked from 'package:readr':  
  
 col\_factor

library(vtable)

Loading required package: kableExtra  
  
Attaching package: 'kableExtra'  
  
The following object is masked from 'package:dplyr':  
  
 group\_rows

winners <- readr::read\_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2023/2023-04-25/winners.csv')

Rows: 163 Columns: 5  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (3): Category, Athlete, Nationality  
dbl (1): Year  
time (1): Time  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

london\_marathon <- readr::read\_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2023/2023-04-25/london\_marathon.csv')

Rows: 42 Columns: 8  
── Column specification ────────────────────────────────────────────────────────  
Delimiter: ","  
chr (1): Official charity  
dbl (6): Year, Applicants, Accepted, Starters, Finishers, Raised  
date (1): Date  
  
ℹ Use `spec()` to retrieve the full column specification for this data.  
ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

#### Preliminary calculations and graphs

#overview of dataset  
vtable(winners)

winners

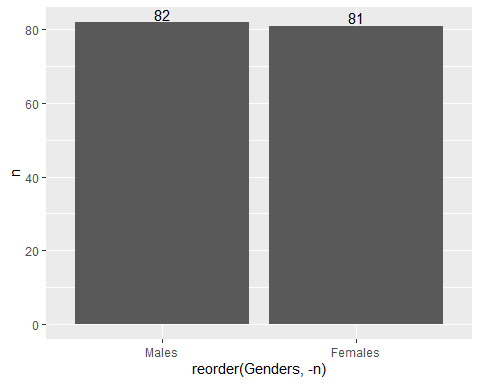
| Name | Class | Values |
| --- | --- | --- |
| Category | character |  |
| Year | numeric | Num: 1981 to 2022 |
| Athlete | character |  |
| Nationality | character |  |
| Time | hms |  |

vtable(london\_marathon)

london\_marathon

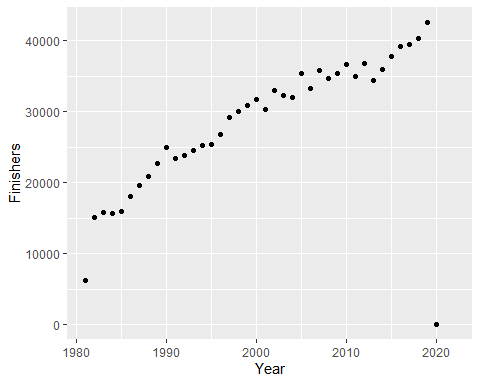
| Name | Class | Values |
| --- | --- | --- |
| Date | Date | Time: 1981-03-29 to 2022-10-02 |
| Year | numeric | Num: 1981 to 2022 |
| Applicants | numeric | Num: 20000 to 457861 |
| Accepted | numeric | Num: 77 to 56398 |
| Starters | numeric | Num: 77 to 42906 |
| Finishers | numeric | Num: 61 to 42549 |
| Raised | numeric | Num: 46.5 to 66.4 |
| Official charity | character |  |

#combine wheelchair men and women with the men and women categories  
groupedGenders <- winners %>%  
 mutate(Genders = if\_else(Category %in% c("Men", "Wheelchair Men"), "Males", "Females"))  
  
#relationship between number of winners and gender  
groupedGenders %>%   
 count(Genders) %>%   
 ggplot(aes(x = reorder(Genders, -n), y = n)) +  
 geom\_col() +  
 geom\_text(aes(label = (n)), vjust = -.2)



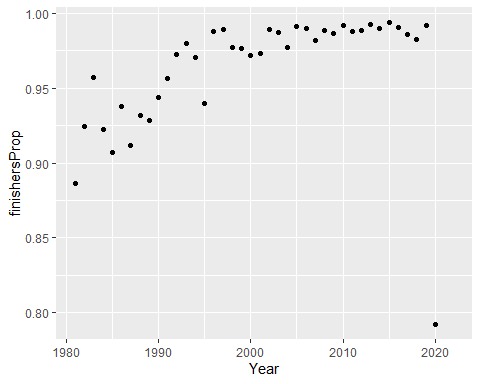
#relationship between number of finishers and year  
london\_marathon %>%  
 ggplot(aes(x = Year, y = Finishers)) +  
 geom\_point()

Warning: Removed 2 rows containing missing values (`geom\_point()`).



#proportion of people finishing the race over the years  
londonMaratonProp <- london\_marathon %>%  
 group\_by(Year) %>%  
 summarise(finishersProp = (Finishers / Starters))   
  
#relationship between years and proportion of finishers  
londonMaratonProp %>%  
 ggplot(aes(x = Year, y = finishersProp)) +  
 geom\_point()

Warning: Removed 2 rows containing missing values (`geom\_point()`).



### Story

The proportion of finishers of the London Marathon has been steadily increasing over time, with the trend approaching a straight line, indicating a more consistent rate of marathon runners are able to finish the race.

### Graph of story

londonMaratonProp %>%  
 ggplot(aes(x = Year, y = finishersProp)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", color = "pink3", se = FALSE) +  
 draw\_text("There were less runners during\nCovid, so this proportion is low.", x = 2015, y = 0.8, size = 5.5, hjust = 0.3, vjust = 0) +  
 scale\_x\_continuous(breaks = seq(1980,2020, by = 5)) +  
 ylim(0.79, 1.005) +  
 labs(y = "Proportion of finishers", x = "Year", title = "Proportion of London Marathon finishers over the years") +  
 theme(text = element\_text(size = 10), panel.grid.minor = element\_blank())

`geom\_smooth()` using formula = 'y ~ x'

Warning: Removed 2 rows containing non-finite values (`stat\_smooth()`).

Warning: Removed 2 rows containing missing values (`geom\_point()`).

