EDA Proposal Statistical

Anant Patel - 0866771

```r
#libraries

library(lubridate)
```

```
Warning: package 'lubridate' was built under R version 4.4.2


Attaching package: 'lubridate'

The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

```r
library(dplyr)
```

```
Warning: package 'dplyr' was built under R version 4.4.2


Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.4.2
```

```r
tuesdata <- tidytuesdayR::tt_load('2023-04-25')
```

```
---- Compiling #TidyTuesday Information for 2023-04-25 ----
--- There are 2 files available ---


-- Downloading files ------------------------------------------------------------

  1 of 2: "winners.csv"
  2 of 2: "london_marathon.csv"
```

```r
tuesdata <- tidytuesdayR::tt_load(2023, week = 17)
```

```
---- Compiling #TidyTuesday Information for 2023-04-25 ----
--- There are 2 files available ---


-- Downloading files ------------------------------------------------------------

  1 of 2: "winners.csv"
  2 of 2: "london_marathon.csv"
```

```r
winners <- tuesdata$winners
london_marathon <- tuesdata$london_marathon
```

```r
View(winners)
View(london_marathon)
```

```r
winners$Time.Seconds <- period_to_seconds(hms(winners$Time))

str(winners)
```

```
spc_tbl_ [163 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Category    : chr [1:163] "Men" "Men" "Men" "Men" ...
 $ Year       : num [1:163] 1981 1981 1982 1983 1984 ...
 $ Athlete    : chr [1:163] "Dick Beardsley (Tie)" "Inge Simonsen (Tie)" "Hugh Jones" "Mike
 $ Nationality : chr [1:163] "United States" "Norway" "United Kingdom" "United Kingdom" ...
 $ Time       : 'hms' num [1:163] 02:11:48 02:11:48 02:09:24 02:09:43 ...
  ..- attr(*, "units")= chr "secs"
 $ Time.Seconds: num [1:163] 7908 7908 7764 7783 7797 ...
 - attr(*, "spec")=
  .. cols(
  ..   Category = col_character(),
  ..   Year = col_double(),
  ..   Athlete = col_character(),
  ..   Nationality = col_character(),
  ..   Time = col_time(format = "")
  .. )
 - attr(*, "problems")=<externalptr>
```

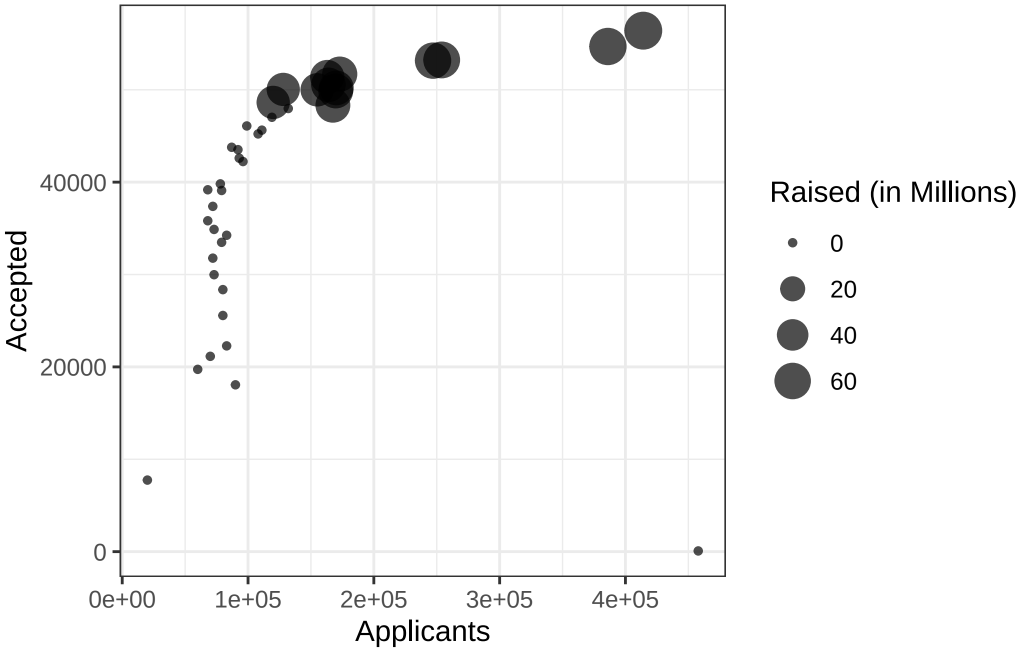
```r
# Factoring the variables
winners$Category <- factor(winners$Category)
winners$Athlete <- factor(winners$Athlete)
winners$Nationality <- factor(winners$Nationality)

str(london_marathon)
```

```
spc_tbl_ [42 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Date            : Date[1:42], format: "1981-03-29" "1982-05-09" ...
 $ Year            : num [1:42] 1981 1982 1983 1984 1985 ...
 $ Applicants      : num [1:42] 20000 90000 60000 70000 83000 80000 80000 73000 72000 73000
 $ Accepted        : num [1:42] 7747 18059 19735 21142 22274 ...
 $ Starters        : num [1:42] 7055 16350 16500 16992 17500 ...
 $ Finishers       : num [1:42] 6255 15116 15793 15675 15873 ...
 $ Raised          : num [1:42] NA NA NA NA NA NA NA NA NA NA ...
 $ Official charity: chr [1:42] NA NA NA NA ...
 - attr(*, "spec")=
  .. cols(
  ..   Date = col_date(format = ""),
  ..   Year = col_double(),
  ..   Applicants = col_double(),
  ..   Accepted = col_double(),
  ..   Starters = col_double(),
  ..   Finishers = col_double(),
  ..   Raised = col_double(),
  ..   `Official charity` = col_character()
  .. )
 - attr(*, "problems")=<externalptr>
```

```
# Handling the NA in Raised
london_marathon$Raised[is.na(london_marathon$Raised)] = 0
```

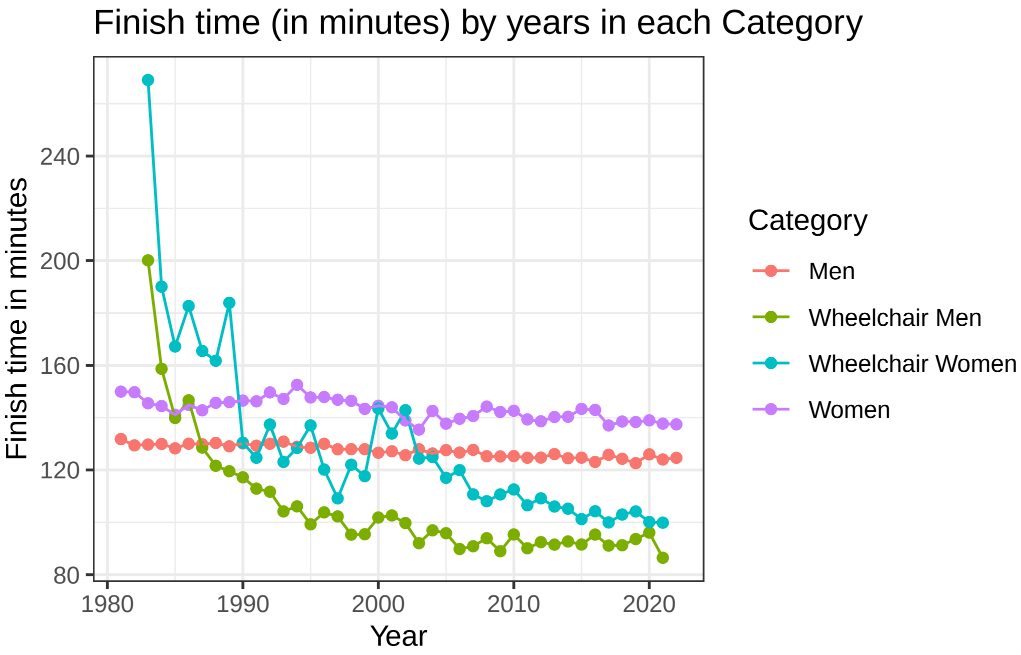
```
options(scip = 999)
# Accepted participants vs finishers by amount raised
london_marathon %>% ggplot(aes(x=Applicants, y = Accepted, size = Raised)) +
  geom_point(alpha = 0.7) +
  scale_size_continuous(name = "Raised (in Millions)")+
  theme_bw()
```

```
Warning: Removed 2 rows containing missing values or values outside the scale range
(`geom_point()`).
```



Question: Does the amount gets raised when the applicants are accepted more? **OR** Claim: when the applicants are accepted more the amount is raised.

```
# Year vs Time by Category
winners %>%
  ggplot(aes(x = Year, y = Time.Seconds / 60, color = Category)) +
  geom_point() +
  geom_line() +
  labs(
    title = "Finish time (in minutes) by years in each Category",
    x = "Year",
    y = "Finish time in minutes"
  ) +
  theme_bw()
```



Wheelchair individuals have some correlation with time to finish the race