EDA Proposal Statistical

Anant Patel - 0866771

Data Preprocessing

The Category, Athlete, and Nationality feature in Winners dataset are then factorized for further analysis.

The data is preprocessed to handle NA values in Raised column in London_marathon dataset by replacing NA values with 0.

Numerical Summaries

```
            Category       Year                      Athlete
Men             :43   Min.   :1981   David Weir          :  8
Wheelchair Men  :39   1st Qu.:1992   Tanni Grey-Thompson :  6
Wheelchair Women:39   Median :2002   David Holding       :  4
Women           :42   Mean   :2002   Eliud Kipchoge      :  4
                      3rd Qu.:2012   Francesca Porcellato:  4
                      Max.   :2022   Ingrid Kristiansen  :  4
                                     (Other)             :133
        Nationality     Time           Time.Seconds
United Kingdom:44   Length:163        Min.   : 5187
Kenya          :30   Class1:hms        1st Qu.: 6550
United States  :11   Class2:difftime   Median : 7675
Switzerland    :10   Mode  :numeric    Mean   : 7608
Ethiopia       : 9                     3rd Qu.: 8418
Norway         : 7                     Max.   :16143
(Other)        :52
```

```
     Date                Year        Applicants        Accepted
Min.   :1981-03-29   Min.   :1981   Min.   : 20000   Min.   :   77
1st Qu.:1991-01-20   1st Qu.:1991   1st Qu.: 78750   1st Qu.:33057
Median :2000-10-18   Median :2000   Median : 94500   Median :43057
```

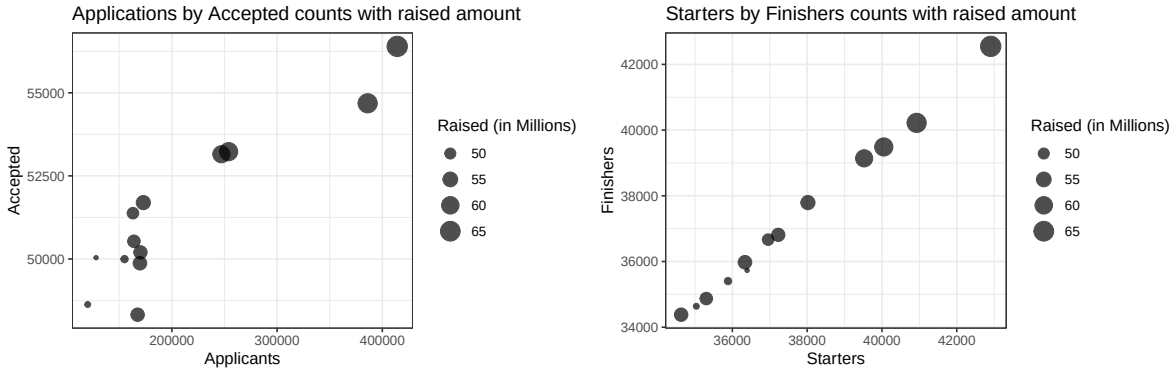
```
 Mean   :2000-10-23   Mean   :2000   Mean   :133354   Mean   :39269
 3rd Qu.:2010-07-23   3rd Qu.:2010   3rd Qu.:163232   3rd Qu.:49903
 Max.   :2020-10-04   Max.   :2020   Max.   :457861   Max.   :56398
    Starters       Finishers         Raised      Official charity
 Min.   :   77   Min.   :   61   Min.   : 0.00   Length:40
 1st Qu.:24488   1st Qu.:23252   1st Qu.: 0.00   Class :character
 Median :31369   Median :30584   Median : 0.00   Mode  :character
 Mean   :28886   Mean   :28145   Mean   :17.67
 3rd Qu.:35671   3rd Qu.:35326   3rd Qu.:48.05
 Max.   :42906   Max.   :42549   Max.   :66.40
```

- The Years of data span from 1981 to 2022 in winners whereas, there is data from 1981 to 2020 on london marathons.
- The highest time to finish a marathon is 16143 seconds which is a outlier.
- In one of the marathons only 77 applicants were accepted and started the marathon.

```
# A tibble: 1 x 8
  Date        Year Applicants Accepted Starters Finishers Raised
  <date>     <dbl>      <dbl>    <dbl>    <dbl>     <dbl>  <dbl>
1 2020-10-04  2020     457861       77       77        61      0
# i 1 more variable: `Official charity` <chr>
```

By subsetting the dataframe, the marathon was conducted in 2020 and there were 4.5 lakhs of applicants but only 77 were accepted.

Visualization



(a) Applicants and Accepted individuals

(b) Starters and Finishers of marathon

Figure 1: Plots showing the relationship of Applicants - Accepted and Finishers - Starters where there is amount of charity is Raised

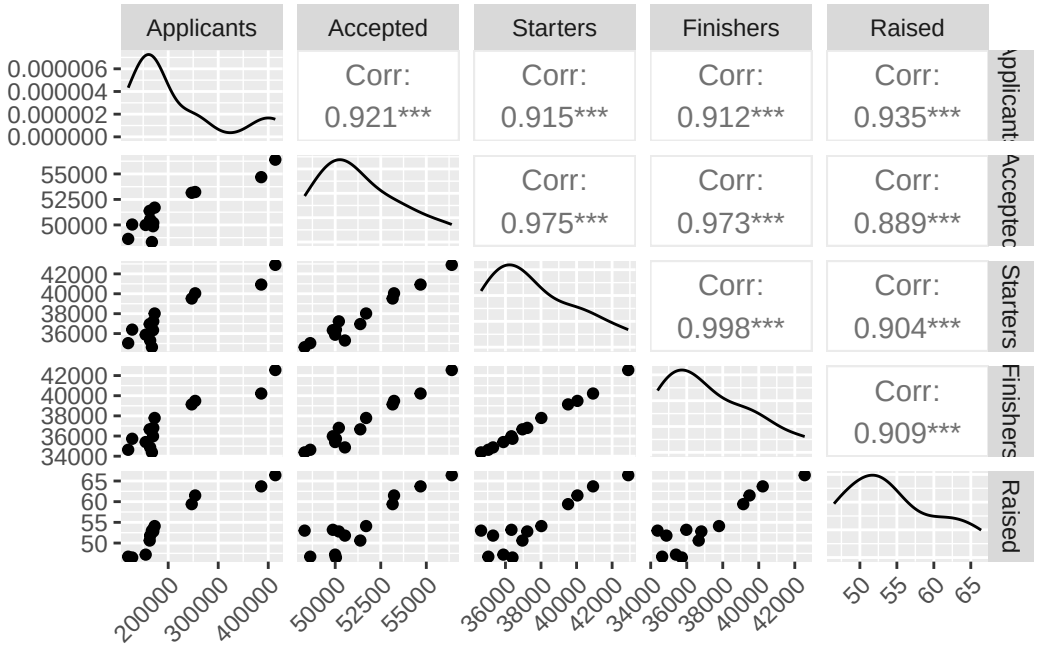


Figure 2: Pair plot between applicants, accepted participants, starters, finishers, and charity amount raised from the marathon

From the scattered bubble plot it is evident that the raised amount increases with the increase in applicants and accepted. Same goes with starters and finishers, there is linearity and the

amount raised also increases.

From the above pair plot it is clear that there is high correlation between Applicants, Accepted, Starters, Finishers, and Raised.

Question: Can we predict the possible charity that can be raised in upcoming london marathons based on the count of Applicants, Accepted participants, starters, and finishers?

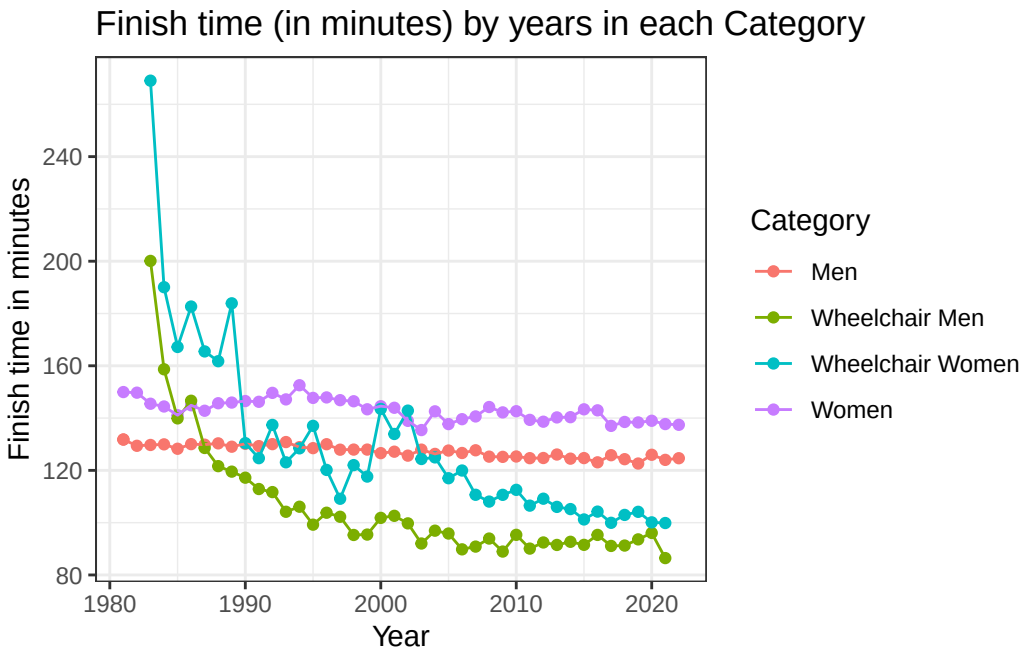


Figure 3: Plot shows the finishing time of each category in each year

Question: Wheelchair individuals have some correlation with time to finish the race? **OR** Does Category have effect on the finish time over the years?

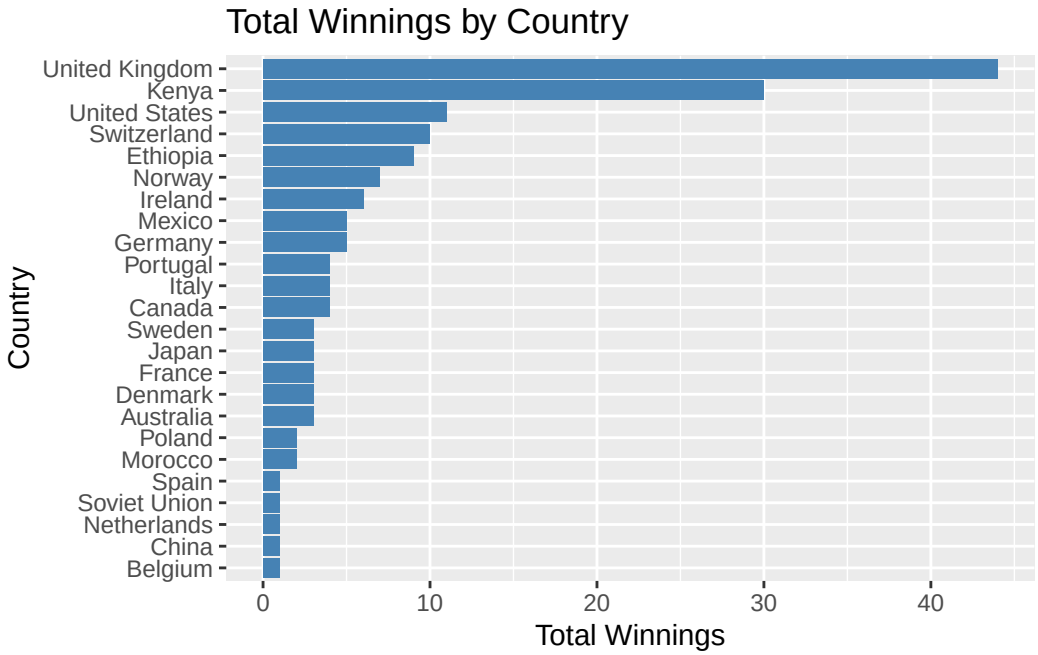


Figure 4: The horizontal barplot show the total winnings of the country in descending order

Question: Does country have significant effect on total winnings?

```r
suppressWarnings(library(dplyr))
suppressWarnings(library(ggplot2))
suppressWarnings(library(lubridate))
suppressWarnings(library(GGally))
#libraries
library(lubridate)
library(dplyr)
library(ggplot2)
library(GGally)
# Loading data
tuesdata <- tidytuesdayR::tt_load('2023-04-25')
tuesdata <- tidytuesdayR::tt_load(2023, week = 17)

winners <- tuesdata$winners
london_marathon <- tuesdata$london_marathon

# Convert the time from hour:min:seconds to seconds
winners$Time.Seconds <- period_to_seconds(hms(winners$Time))

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

# Handling the NA in Raised
london_marathon$Raised[is.na(london_marathon$Raised)] = 0
london_marathon <- london_marathon[rowSums(is.na(london_marathon)) <= 2,]
summary(winners)
summary(london_marathon)
# Subset the dataframe for starters = 77
london_marathon[london_marathon$Starters == 77,]
# Option ot print values without scientific notation
options(scipen = 999)
# Applicants vs Accepted participants by amount raised
london_marathon %>%
  filter(Raised > 0 ) %>%
  ggplot(aes(x=Applicants, y = Accepted, size = Raised)) +
  geom_point(alpha = 0.7) +
  scale_size_continuous(name = "Raised (in Millions)")+
  labs(title="Applications by Accepted counts with raised amount") +
  theme_bw()
```

```r
# Starters and finishers by amount raised
london_marathon %>%
  filter(Raised > 0 ) %>%
  ggplot(aes(x=Starters, y = Finishers, size = Raised)) +
  geom_point(alpha = 0.7) +
  scale_size_continuous(name = "Raised (in Millions)")+
  labs(title="Starters by Finishers counts with raised amount") +
  theme_bw()

# ggpairs plot
london_marathon[,c("Applicants","Accepted","Starters","Finishers","Raised")] %>%
  filter(Raised > 0) %>%
  ggpairs() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # chat gpt helped me
# 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()

winners_count <- data.frame(table(winners$Athlete))
names(winners_count) <- c("Athlete","Frequency")
#
winners_nationality <- unique(left_join(winners_count,
                                        winners[,c("Athlete","Nationality")],
                                        by="Athlete"))
grouped_nationality <- winners_nationality %>%
  group_by(Nationality) %>%
  summarise(Total_winnings = sum(Frequency))

grouped_nationality %>%
  ggplot(aes(x=reorder(Nationality,Total_winnings), y=Total_winnings)) +
  geom_bar(stat="identity",fill="steelblue") +
  labs(title="Total Winnings by Country", x = "Country", y = "Total Winnings")+
  coord_flip()
```