EDA Proposal Statistical

Anant Patel - 0866771

```
#libraries

library(lubridate)
```

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

```
library(dplyr)
```

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

```
library(ggplot2)
```

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

```
library(GGally)
```

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

```
tuesdata <- tidytuesdayR::tt_load('2023-04-25')
tuesdata <- tidytuesdayR::tt_load(2023, week = 17)

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

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

# str(winners)

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

# str(london_marathon)

# Handling the NA in Raised
london_marathon$Raised[is.na(london_marathon$Raised)] = 0
london_marathon <- london_marathon[rowSums(is.na(london_marathon)) <= 2,]
```

```r
summary(winners)
```

```
             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
```

```r
summary(london_marathon)
```

```
      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.

```r
london_marathon[london_marathon$Starters == 77,]
```

```
# 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>
```

```r
options(scipen = 999)
# Accepted participants vs finishers 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()
```

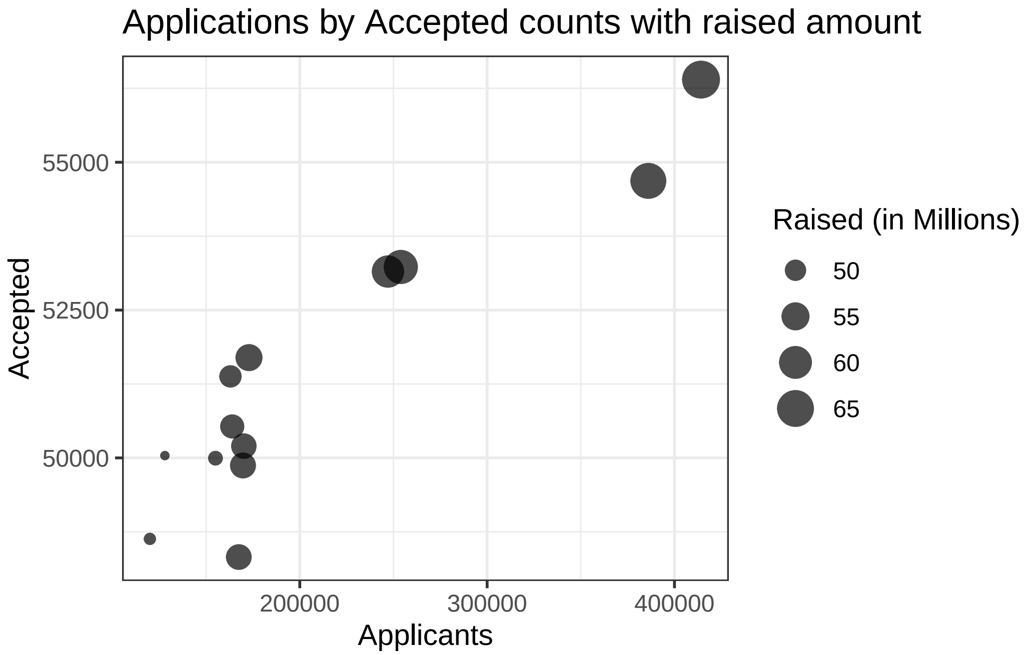
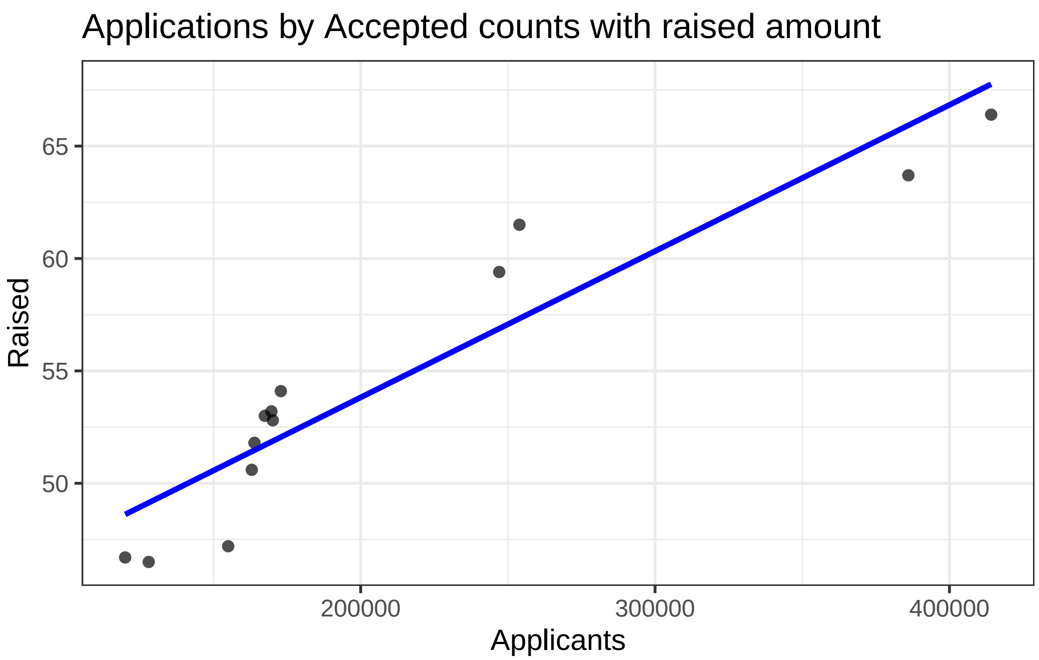


Figure 1: Scatter plot for total Applicants vs. Accepted applications and funds raised by each race

Question: Does the amount gets raised when the applicants are accepted more? Proposed Solution: Fitting a linear model can help.

```
lmod <- lm(Raised ~ Applicants + Accepted + Starters + Finishers,london_marathon)

# lmod <- lm(Raised ~ Applicants,london_marathon[london_marathon$Raised > 0,])

london_marathon %>%
  filter(Raised > 0) %>%
  # filter(Raised > 0 & Applicants < 450000) %>%
  filter(Applicants < 450000) %>%
  ggplot(aes(x=Applicants, y = Raised)) +
  geom_point(alpha = 0.7) +
  scale_size_continuous(name = "Raised (in Millions)")+
  labs(title="Applications by Accepted counts with raised amount") +
  theme_bw() +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE, color = "blue")
```



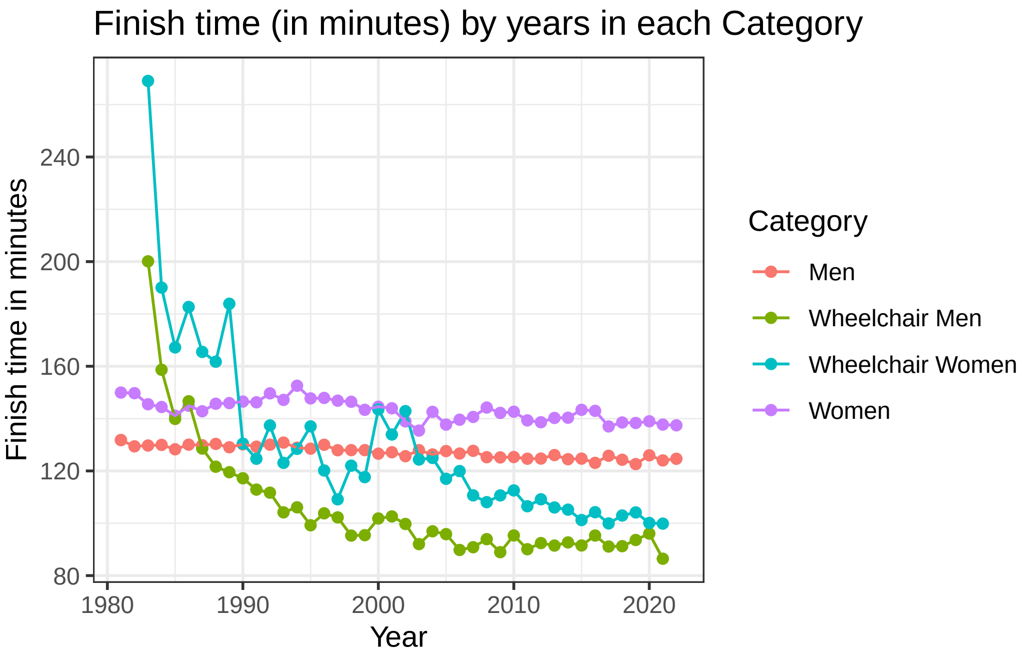
Applications by Accepted counts with raised amount

```
summary(lmod)
```

```
Call:
lm(formula = Raised ~ Applicants + Accepted + Starters + Finishers,
    data = london_marathon)

Residuals:
     Min       1Q   Median       3Q      Max
-30.6615 -12.6324  -0.8866  10.7487  26.8530

Coefficients:
                Estimate   Std. Error t value Pr(>|t|)
(Intercept) -40.94549533  12.37057088  -3.310 0.002171 **
Applicants    0.00011584   0.00003147   3.681 0.000777 ***
Accepted     -0.00192098   0.00196776  -0.976 0.335653
Starters      0.00046931   0.00631594   0.074 0.941190
Finishers     0.00373246   0.00673803   0.554 0.583143
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.8 on 35 degrees of freedom
```

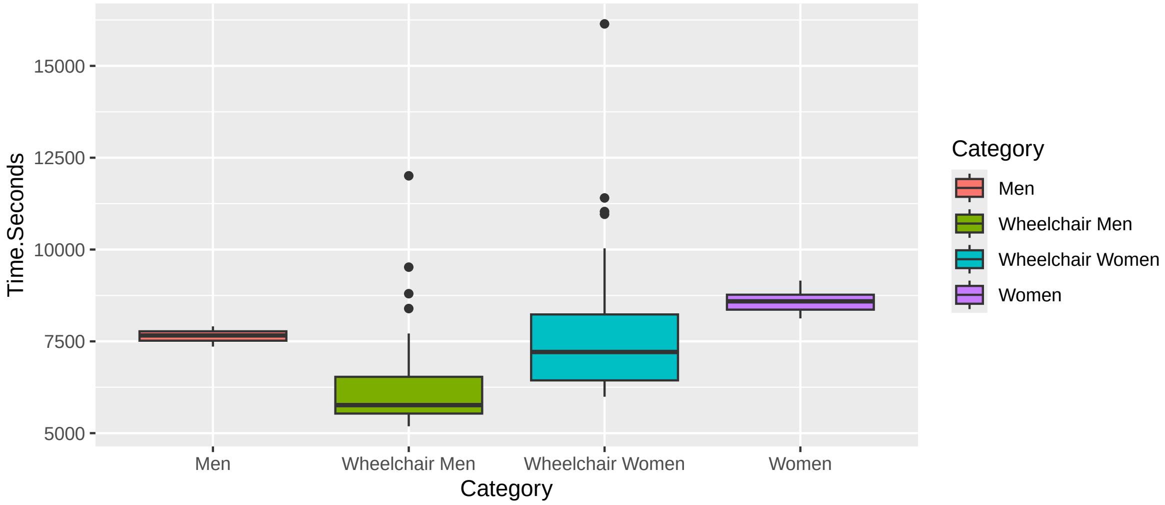
```
Multiple R-squared:  0.6696,    Adjusted R-squared:  0.6319
F-statistic: 17.73 on 4 and 35 DF,  p-value: 0.00000004866
```

```
# 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()
```



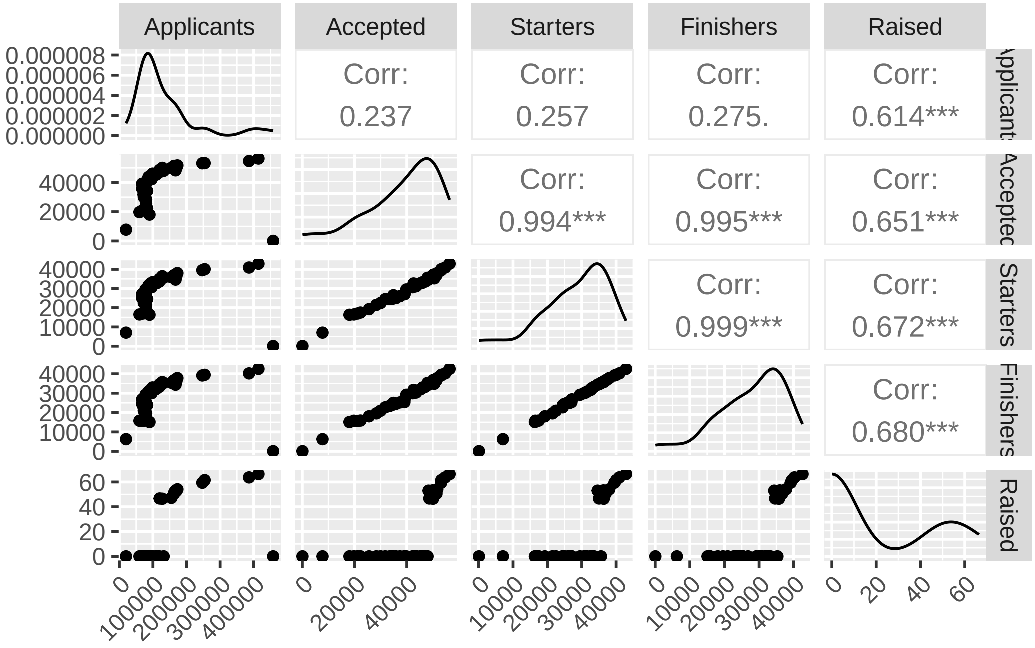
Question: Wheelchair individuals have some correlation with time to finish the race? Proposed Solution: ANOVA test to identify the relationship between Category and time to finish race

```r
winners %>%
  ggplot(aes(x = Category, y = Time.Seconds, fill = Category)) +
  geom_boxplot()
```



Question: Is Womens' time normally distributed? Proposed Solution: Histogram, Boxplot, qqplot, and shapiro-wilks

```r
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
```

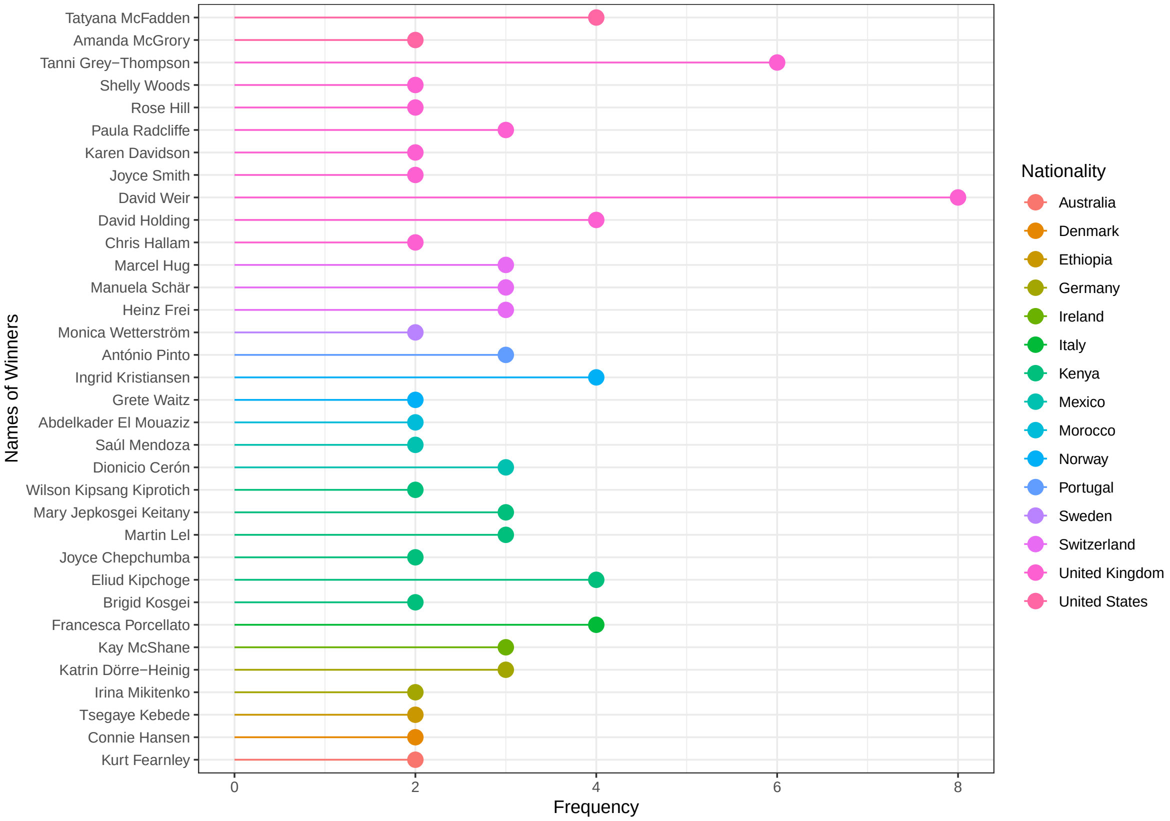


OpenAI. (2024). ChatGPT [Large language model]. https://chatgpt.com

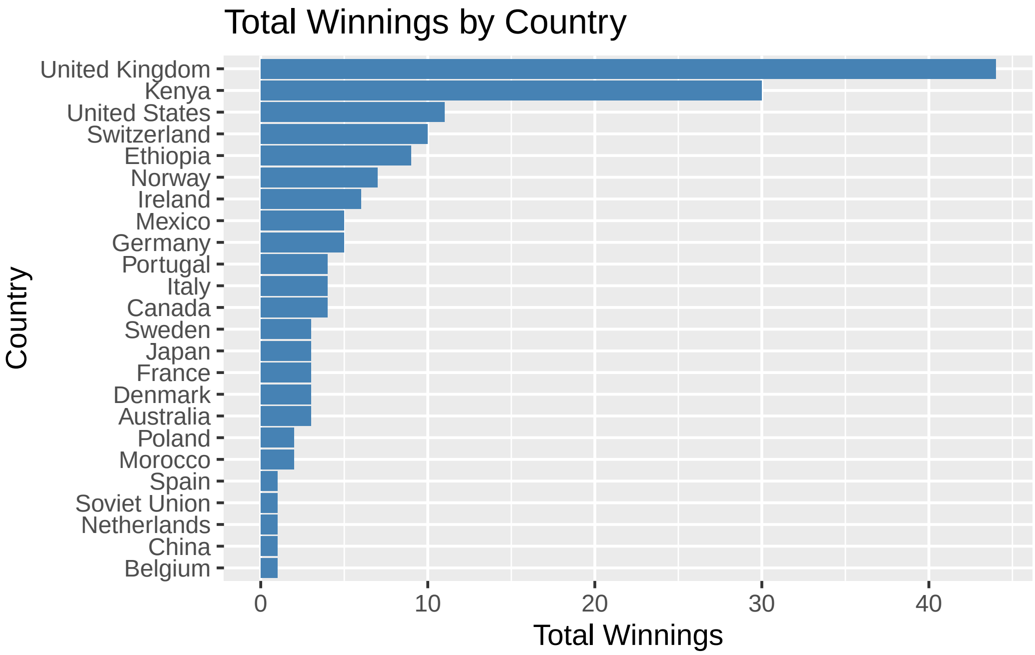
(OpenAI, 2024)

Question: Can we predict the Raised amount based on the Applicants, Accepted, Starters and Finishers. proposed solution: Multiple Linear Regression

```
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"))

winners_nationality[winners_nationality$Frequency > 1,] %>%
  arrange(Nationality) %>%
  mutate(name=factor(Athlete, levels=Athlete)) %>%
  ggplot( aes(x=name, y=Frequency, color=Nationality)) +
  geom_segment( aes(x=name, xend=name, y=Frequency,yend=0)) +
  geom_point( size=4) +
  coord_flip() +
  theme_bw() +
  xlab("Names of Winners")
```

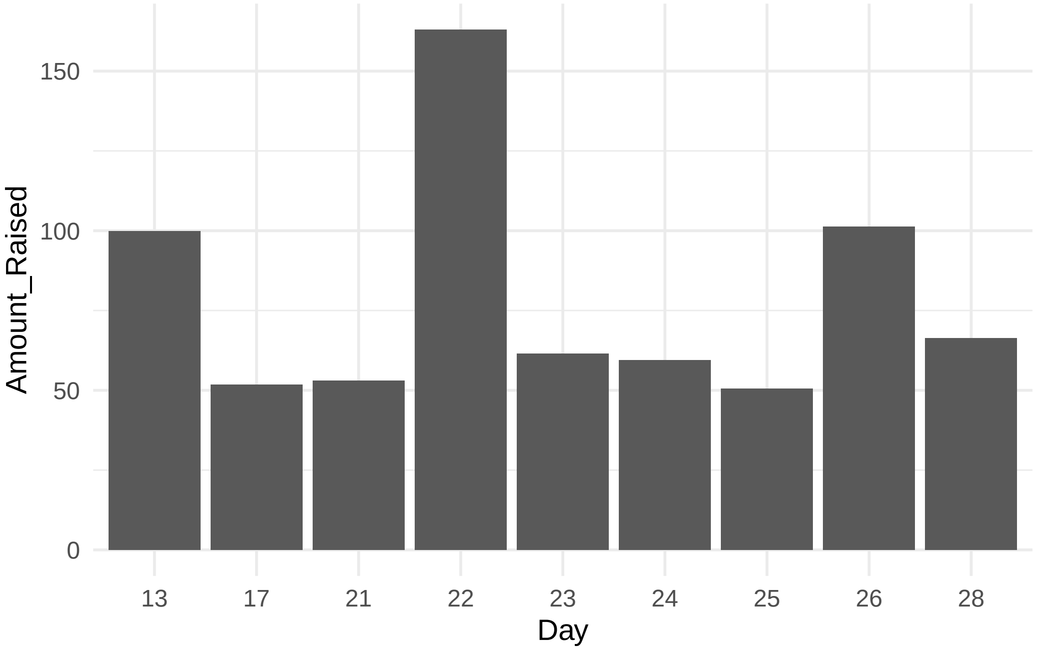


```
grouped_nationality <- winners_nationality %>%
  group_by(Nationality) %>%
  summarise(Total_winnings = sum(Frequency))

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()
```



Question: Does country have significant effect on total winnings? Proposed Solution: ANOVA

```r
raised_mday <- data.frame(table(mday(london_marathon$Date), london_marathon$Raised))
raised_mday$Var2 <- as.numeric(as.character(raised_mday$Var2))
names(raised_mday) <- c("Day","Amount_Raised","Frequency")

raised_mday %>%
  filter(Amount_Raised > 0 & Frequency > 0) %>%
  ggplot(aes(x=Day, y=Amount_Raised)) +
  geom_col() +
  theme_minimal()
```



Question: Does the Day of month have significant effect on amount raised Proposed Solution: ANOVA