London Marathon Data Analysis

Introduction

The London Marathon, or TCS London Marathon, unites people from all walks of life. Conceived in 1981 by Chris Brasher and John Disley, it's the UK's second-largest annual road race. Originally in April, it temporarily moved to October in 2020-2022 due to COVID-19. The flat, scenic course runs along the River Thames, starting at Blackheath and finishing at The Mall.

Beyond being a race, it's a diverse experience. The mass race welcomes all fitness levels, elite long-distance runners compete, and wheelchair races showcase top-level competition for both men and women.

Since 1981, the marathon has raised over £1 billion for charities. In 2019, a record-breaking £66.4 million was raised, marking it as the highest single-day fundraising event in its history.

How was the information collected?

Data seems sourced from official records, the London Marathon's website, event organizers, media outlets like BBC, and reliable sources associated with the event, compiling information from participant registrations, event statistics, official reports, and media coverage over the years.

For more details:

https://en.wikipedia.org/wiki/London_Marathon

Now, let's dive into the details of the marathon with the comprehensive data:

The "LondonMarathon" package includes two datasets:

London Marathon Data:

Date, Year: Date and corresponding year of the London Marathon.

Applicants, Accepted: Count of applicants and accepted participants.

Starters, Finishers: Participants who started and successfully completed the marathon.

Raised: Total raised amount during the marathon.

Official Charity: Designated charity for the year.

Winners Data:

Category, Year: Winner's category and the year of the London Marathon.

Athlete, Nationality: Winning athlete's name and nationality.

Time: Duration taken by the winning athlete to complete the marathon.

Cases: 42 in "London Marathon Data," 165 in "Winners Data."

Winners Data

Category	Year	Athlete	Nationality	Time
Men	1981	Dick Beardsley (Tie)	United States	0.09152778
Men	1981	Inge Simonsen (Tie)	Norway	0.09152778
Men	1982	Hugh Jones	United Kingdom	0.08986111
Men	1983	Mike Gratton	United Kingdom	0.09008102
Men	1984	Charlie Spedding	United Kingdom	0.09024306
Men	1985	Steve Jones	United Kingdom	0.08907407

London Marathon Data

Date	Year	Applicants	Accepted	Starters	Finishers	Raised	Official charity
1981-03-29	1981	20000	7747	7055	6255	NA	NA
1982-05-09	1982	90000	18059	16350	15116	NA	NA
1983-04-17	1983	60000	19735	16500	15793	NA	NA
1984-05-13	1984	70000	21142	16992	15675	NA	NA
1985-04-21	1985	83000	22274	17500	15873	NA	NA
1986-04-20	1986	80000	25566	19261	18067	NA	British Sports Association for the Disabled (autistic)

Data Wrangling

Overview of Data

```
##    Category              Year        Athlete          Nationality
##  Length:165         Min.   :1981   Length:165         Length:165
##  Class :character   1st Qu.:1992   Class :character   Class :character
##  Mode  :character   Median :2002   Mode  :character   Mode  :character
##                     Mean   :2002
##                     3rd Qu.:2012
##                     Max.   :2023
##       Time
##  Min.   :0.06003
##  1st Qu.:0.07581
##  Median :0.08883
##  Mean   :0.08808
##  3rd Qu.:0.09740
##  Max.   :0.18684
```

The dataset "winners" is free of missing values, and all variables possess suitable data types.

The "Category" column within the Winners dataset consists of four distinct categories: Men, Women, Wheelchair Men, and Wheelchair Women.

```
##
##
## |Category         |
## |:----------------|
## |Men              |
## |Women            |
## |Wheelchair Men   |
## |Wheelchair Women |
```

Do variables have appropriate type?

Here is the data types summary.

Table 3: Winners Data

Column	Class	Description	Example
Category	character	Category of race	Men
Year	integer	Year	1981
Athlete	character	Name of the winner	Dick Beardsley(Tie)
Nationality	character	Nationality of the winner	United States
Time	character	Winning time	02:11:48

Table 4: London Marathon Data

Column	Class	Description	Example
Date	character	Date of the race	1981-03-29
Year	integer	Year	1981
Applicants	integer	Number of people who applied	20000
Accepted	integer	Number of people accepted	7747
Starters	integer	Number of people who started	7055
Finishers	integer	Number of people who finished	6255
Raised	integer	Amount raised for charity(£ millions)	46.5
Official.charity	character	Official charity	SportsAid

Are there any strange aspects to the data?

In the london_marathon dataset, observation of number of applicants in year 1981-2012 are in a unit of 1000. And observations of number of applicants, accepted, starters, and finishers in year 2020 are abnormally low.

Winners Data - Analysis & Conclusions

Analysis of Winners Dataset

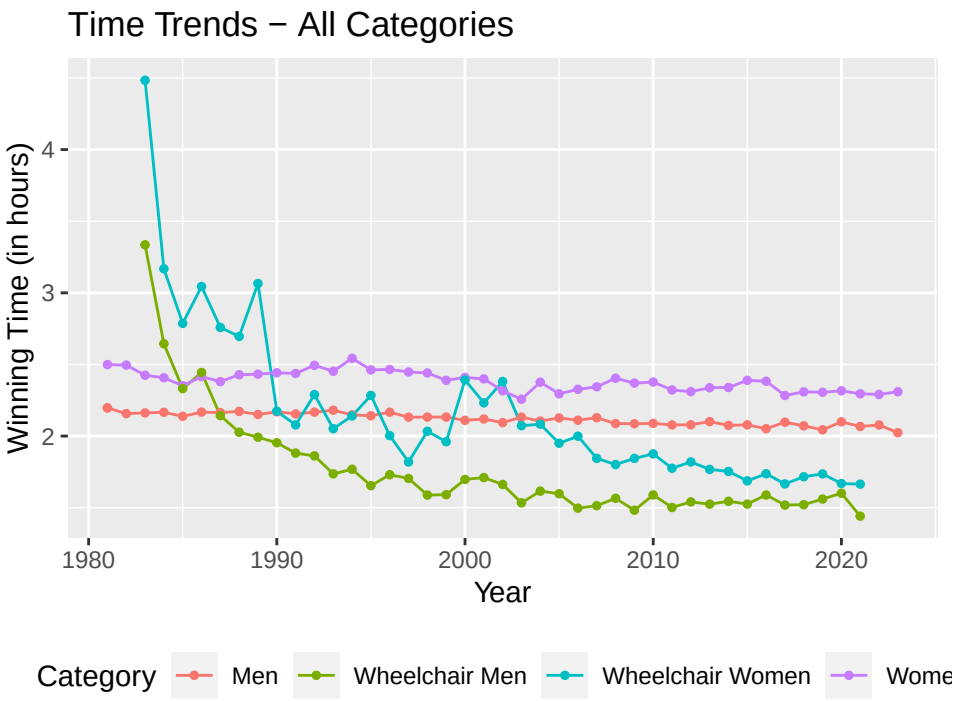


Figure 1: Time trends in winning times for different categories over the years.

Trends in winning times over the year

Athletic performance in the Men and Women categories has shown a consistent improvement, as reflected in decreasing winning times from approximately 0.09 to 0.08 for Men and 0.104 to 0.096 for Women. Wheelchair Men and Wheelchair Women have also experienced significant performance enhancements, with winning times dropping from 0.1389 to 0.06 and 0.1868 to 0.069, respectively. Overall, there is a notable trend of improved performance across all categories, marked by substantial drops in winning times over the years.

Compare Winning Times Between Different Categories (Men vs. Women, Wheelchair Men vs. Wheelchair Women)

t-test for Men vs. Women

```
##
##  Welch Two Sample t-test
##
## data:  Time by Category
## t = -21.25, df = 69.816, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Men and group Women is not equal to 0
## 95 percent confidence interval:
##  -0.2865984 -0.2374147
## sample estimates:
##   mean in group Men mean in group Women
##            2.121004            2.383010
```

The comparison of average finishing times between Men and Women in the marathon using the Welch Two Sample t-test shows a highly significant result (t-value = -21.25). This means Men, on average, finish the marathon faster than Women. We accounted for differences in variances using Welch's correction, resulting in approximately 69.816 degrees of freedom.

The p-value is incredibly small (< 2.2e-16), strongly indicating that the mean winning time for Men is significantly different from that of Women. The 95% confidence interval (-0.2865984 to -0.2374147) supports this, as it doesn't include zero.

The evidence suggests a notable difference, favoring Men, and the average winning time for Men (2.121004) is significantly lower than for Women (2.383010). So, we have strong statistical grounds to reject the idea that there's no difference, confirming that Men tend to finish the marathon faster than Women.

t-test for Wheelchair Men vs. Wheelchair Women

```
##
##  Welch Two Sample t-test
##
## data:  Time by Category
## t = -3.684, df = 66.574, p-value = 0.0004625
## alternative hypothesis: true difference in means between group Wheelchair Men and group Wheelchair W
## 95 percent confidence interval:
##  -0.6161296 -0.1830727
## sample estimates:
##   mean in group Wheelchair Men mean in group Wheelchair Women
##                       1.761944                       2.161546
```

The t-test results for Wheelchair Men vs. Wheelchair Women reveal a significant difference in mean winning times. The negative t-value (-3.684) suggests that, on average, Wheelchair Women have a higher winning time. The degrees of freedom (df) are 66.574, calculated using Welch's method.

The p-value of 0.0004625 is less than 0.05, providing strong evidence to reject the null hypothesis of equal mean winning times. The 95% confidence interval (-0.6161296 to -0.1830727) indicates that the mean winning time for Wheelchair Women is likely to be between 0.1831 and 0.6161 lower than that of Wheelchair Men.

Overall, there is a statistically significant and lower mean winning time for Wheelchair Women compared to Wheelchair Men.

ANOVA for Men vs. Women vs. Wheelchair Men vs. Wheelchair Women

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Category      3  8.036   2.679   24.34 4.81e-13 ***
## Residuals   161 17.714   0.110
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis (H0): There is no significant difference in mean winning times among the categories. Alternative Hypothesis (H1): There is a significant difference in mean winning times among at least one pair of categories.

I reject the null hypothesis, concluding that there are significant differences in mean winning times across at least one pair of categories (Men, Women, Wheelchair Men, Wheelchair Women). F-Test: The F value of 24.34 is higher than expected by chance, signifying significant differences in mean winning times among categories.

p-value: With a tiny p-value (4.81e-13), there is substantial evidence against the null hypothesis, implying significant differences in mean winning times.

Holder's of the Current Records

Category	Year	Athlete	Nationality	Time
Men	2023	Kelvin Kiptum	Kenya	0.08431713
Women	2003	Paula Radcliffe	United Kingdom	0.09403935
Wheelchair Men	2021	Marcel Hug	Switzerland	0.06003472

Category	Year	Athlete	Nationality	Time
Wheelchair Women	2021	Manuela Schär	Switzerland	0.06935185

Winners Count by Category and Nationality

Category	Nationality	count
Men	Kenya	17
Wheelchair Men	United Kingdom	16
Wheelchair Women	United Kingdom	15
Women	Kenya	14

Marathon Dataset - Analysis & Conclusions

When analyzing the London Marathon Data, it was observed that there are instances where certain data points are missing. Specifically, there are 43 occurrences where the information is not available.

```
## Missing Value Count: 43
```

Percentage Of Missing Values In London Marathon

The London Marathon dataset has missing values, especially in the “Raised” and “Official charity” columns. If “Official charity” is NULL, “Raised” is considered 0.

If “Official charity” is present and “Raised” is missing, the value can be filled with the average of other data. Additionally, missing values in other fields are consistent and likely due to the absence of data for the last two years, as per official documentation on Github.

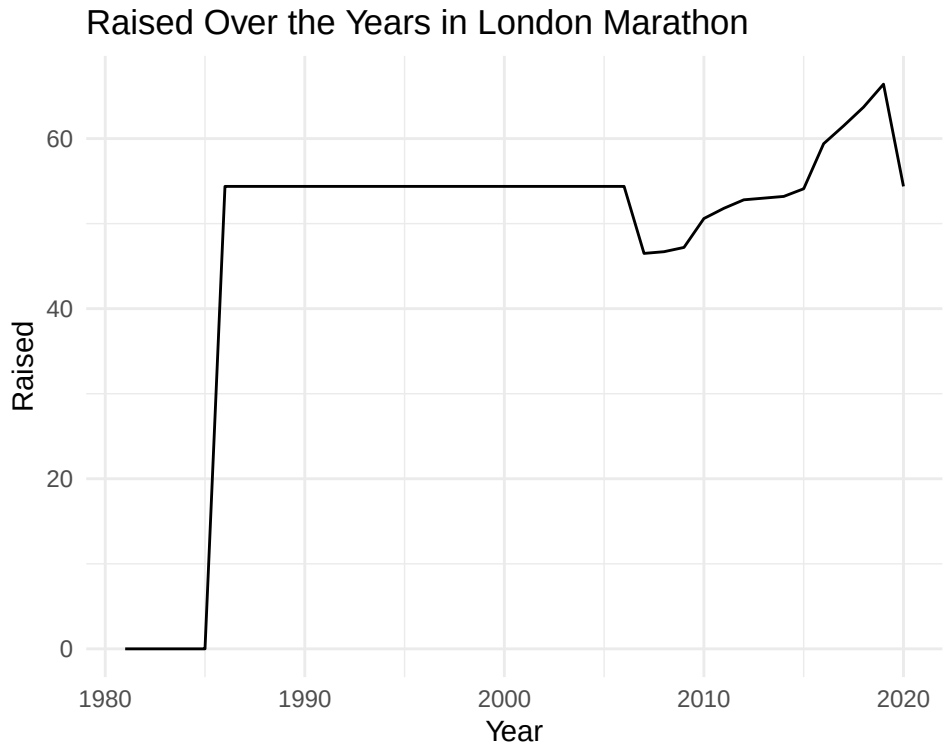
It’s advisable to exclude entries from these two years in the analysis.

Data Cleaning and Imputation for London Marathon Dataset

After performing thorough data cleaning and imputation processes on the London Marathon Dataset, it is noteworthy that the dataset is now free from missing values, and the count of missing values has been reduced to 0.

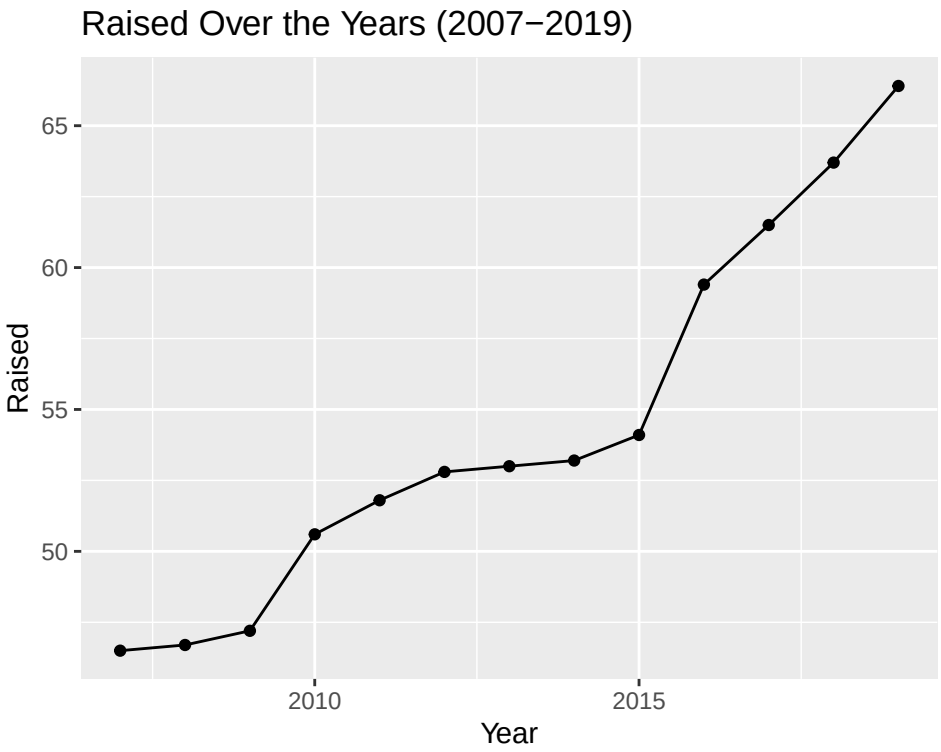
```
## New Missing Value Count: 0
```

Filled Raised Data Chart

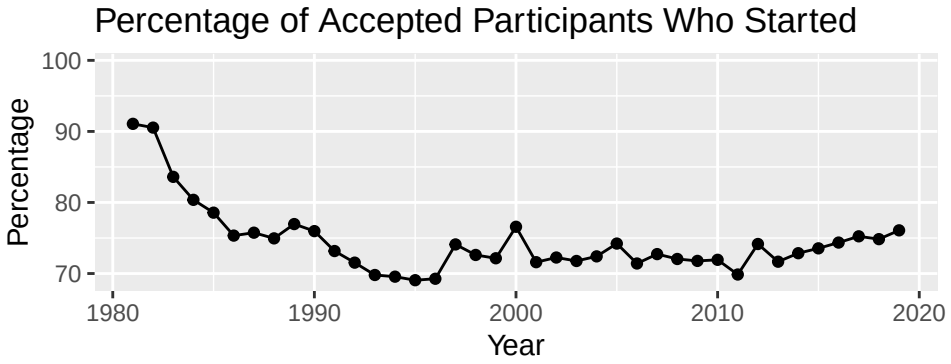


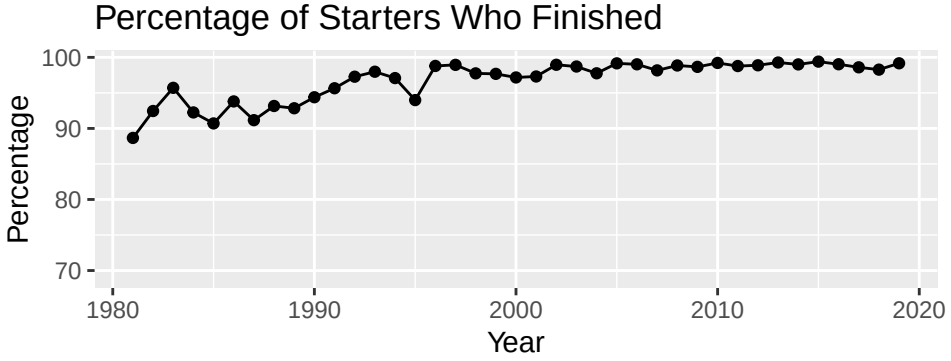
The filled Raised data in the line chart doesn't provide useful insights for analyzing the changes. To better focus on this aspect, **we should narrow our attention to the data spanning from 2007 to 2019.**

Filtered Line Chart: Changes in Raised Data (2007-2019)



Calculate The Percentage Of Accepted Participants Who Started Race And The Percentage Of Starters Who Finished.





Accepted to Started:

- Over the years, there is a general trend of a decline in the percentage of accepted participants who actually start the race.

Started to Finished:

- The percentage of starters who finish the race seems to be relatively stable, generally ranging from the high 80s to mid-90s.

Notable Observation (2020):

- In 2020, there's a notable observation in the "Accepted to Started" column. The percentage is 100%, indicating that all accepted participants started the race. However, the "Started to Finished" percentage is lower (79.22%), indicating that not all starters completed the race.
- This discrepancy could be due to the impact of external factors such as the COVID-19 pandemic, leading to a different pattern in that particular year.

Exploring Distribution Of Winners By Nationality In London Marathon Data

Nationality	Count
United Kingdom	44
Kenya	31
United States	11

The United Kingdom, Kenya, and the United States stand out as the top three countries with the highest counts of both runners and winners.