R_Project_Analysis

Akshay Suresh Varma

2023-12-4

Loading Packages

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.2

## Warning: package 'readr' was built under R version 4.3.2

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.3.2

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(mosaicData)
```

```
## Warning: package 'mosaicData' was built under R version 4.3.2
```

```
remotes::install_github("nrennie/LondonMarathon")
```

```
## Skipping install of 'LondonMarathon' from a github remote, the SHA1 (c83c6806) has not changed since
##   Use `force = TRUE` to force installation
```

```
data(winners, package = "LondonMarathon")
data(london_marathon, package = "LondonMarathon")
```

```
winners_data <- winners
winners_data
```

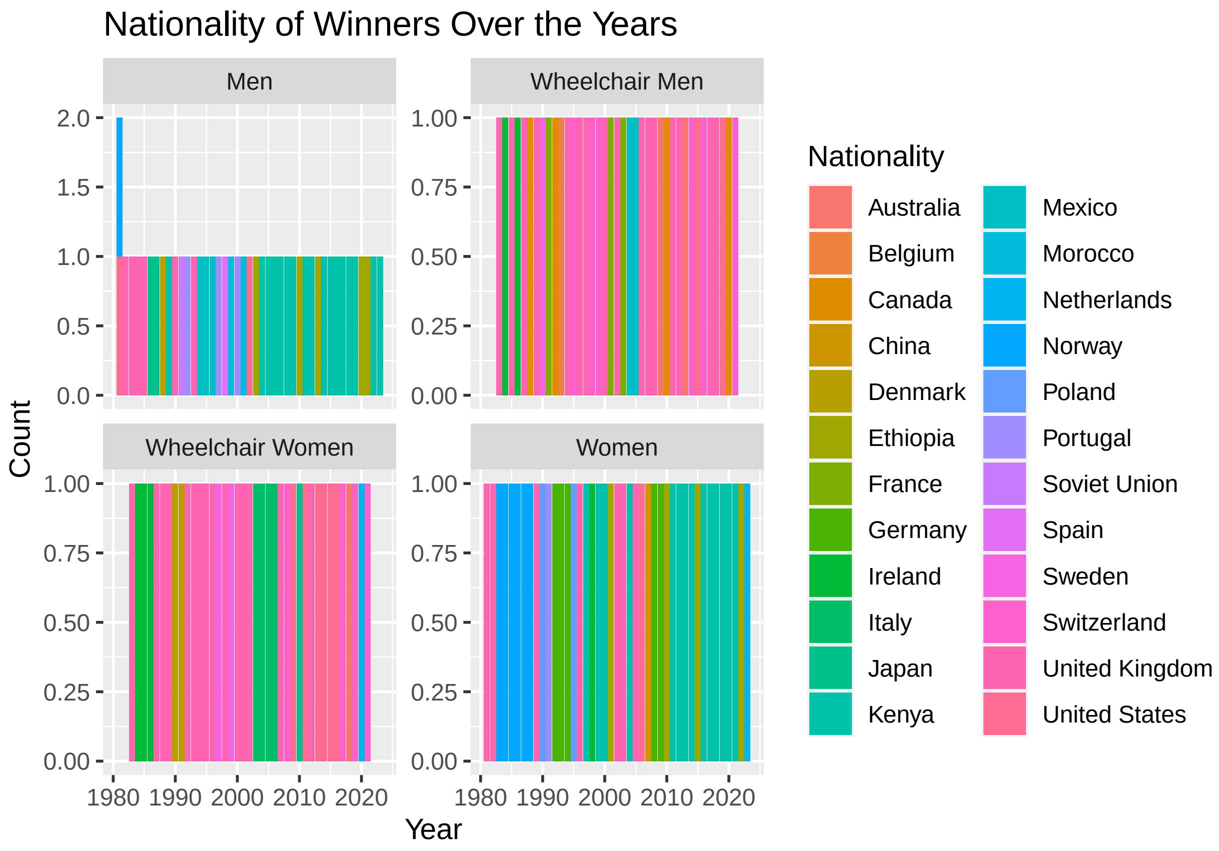
```
## # A tibble: 165 x 5
##    Category  Year Athlete              Nationality    Time
##    <chr>    <dbl> <chr>                <chr>          <times>
##  1 Men       1981 Dick Beardsley (Tie) United States  0.09152778
##  2 Men       1981 Inge Simonsen (Tie)  Norway         0.09152778
##  3 Men       1982 Hugh Jones           United Kingdom 0.08986111
##  4 Men       1983 Mike Gratton         United Kingdom 0.09008102
##  5 Men       1984 Charlie Spedding     United Kingdom 0.09024306
##  6 Men       1985 Steve Jones          United Kingdom 0.08907407
##  7 Men       1986 Toshihiko Seko       Japan          0.09030093
##  8 Men       1987 Hiromi Taniguchi     Japan          0.09016204
##  9 Men       1988 Henrik Jørgensen     Denmark        0.09050926
## 10 Men       1989 Douglas Wakiihuri    Kenya          0.08961806
## # i 155 more rows
```

```
marathon_data <- london_marathon
marathon_data
```

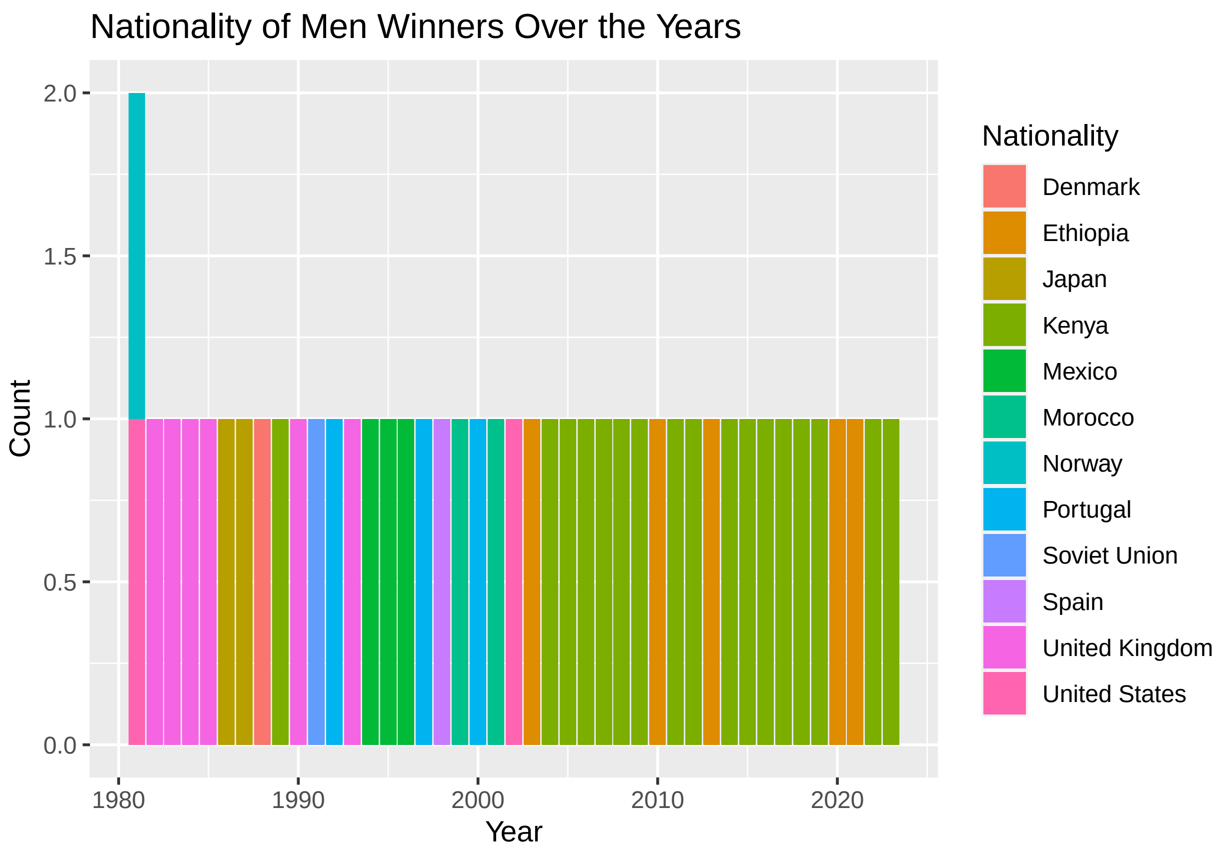
```
## # A tibble: 42 x 8
##    Date        Year Applicants Accepted Starters Finishers Raised
##    <date>     <dbl>      <dbl>    <dbl>    <dbl>     <dbl>  <dbl>
##  1 1981-03-29  1981      20000     7747     7055      6255     NA
##  2 1982-05-09  1982      90000    18059    16350     15116     NA
##  3 1983-04-17  1983      60000    19735    16500     15793     NA
##  4 1984-05-13  1984      70000    21142    16992     15675     NA
##  5 1985-04-21  1985      83000    22274    17500     15873     NA
##  6 1986-04-20  1986      80000    25566    19261     18067     NA
##  7 1987-05-10  1987      80000    28364    21485     19586     NA
##  8 1988-04-17  1988      73000    29979    22469     20932     NA
##  9 1989-04-23  1989      72000    31772    24452     22701     NA
## 10 1990-04-22  1990      73000    34882    26500     25013     NA
## # i 32 more rows
## # i 1 more variable: `Official charity` <chr>
```

```
library(ggplot2)

ggplot(winners_data, aes(x = Year, y = ..count.., fill = Nationality)) +
  geom_bar(position = "stack") +
  labs(title = "Nationality of Winners Over the Years",
       x = "Year",
       y = "Count",
       fill = "Nationality") +
  facet_wrap(~Category, scales = "free_y")
```

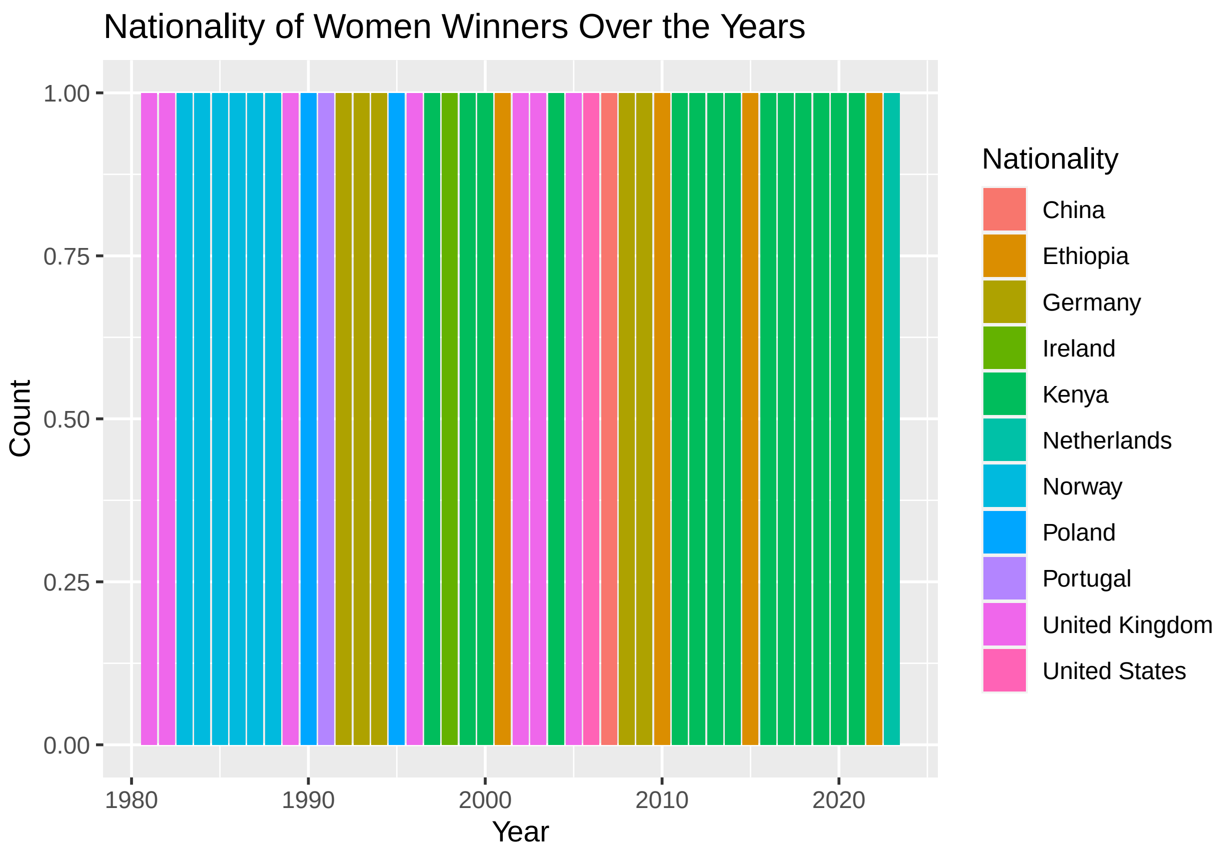
```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



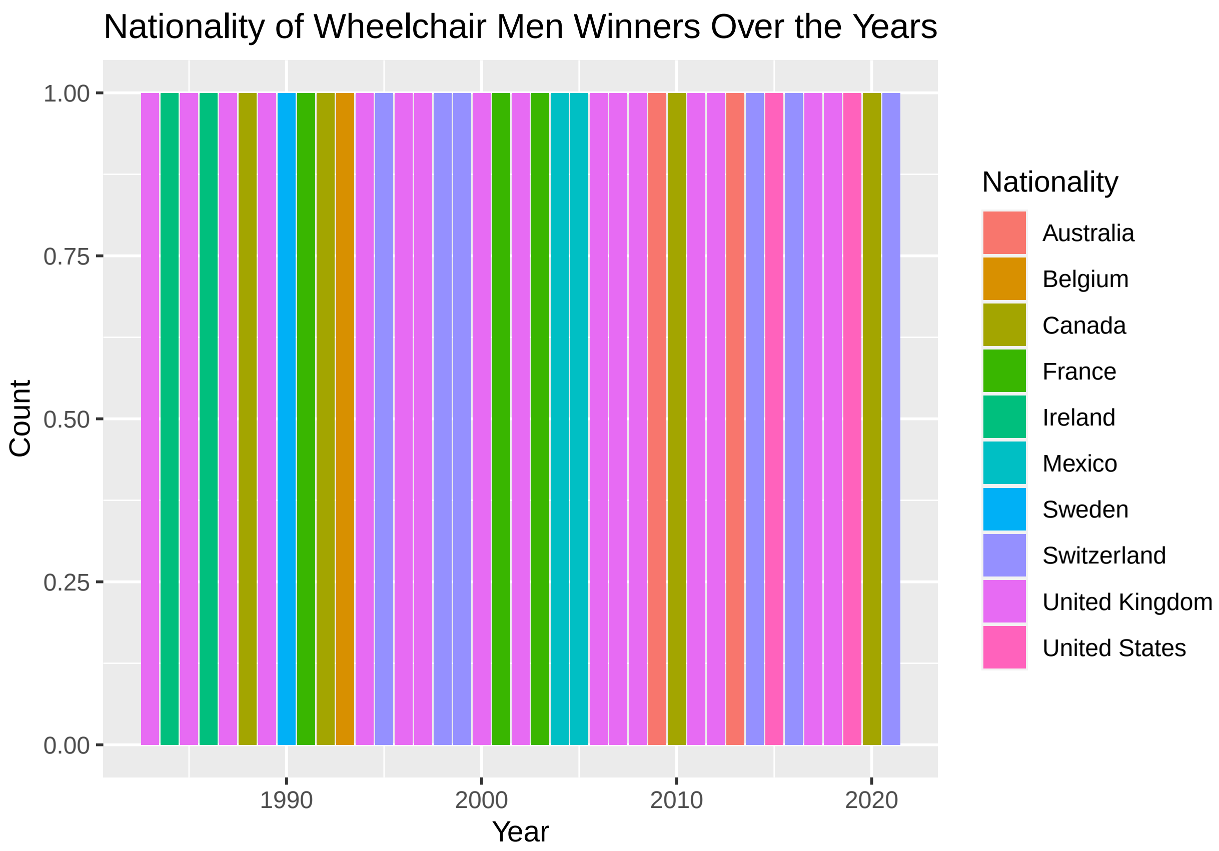
```
library(ggplot2)

ggplot(winners_data[winners_data$Category == "Men", ], aes(x = Year, fill = Nationality)) +
  geom_bar(position = "stack") +
  labs(title = "Nationality of Men Winners Over the Years",
       x = "Year",
       y = "Count",
       fill = "Nationality")
```



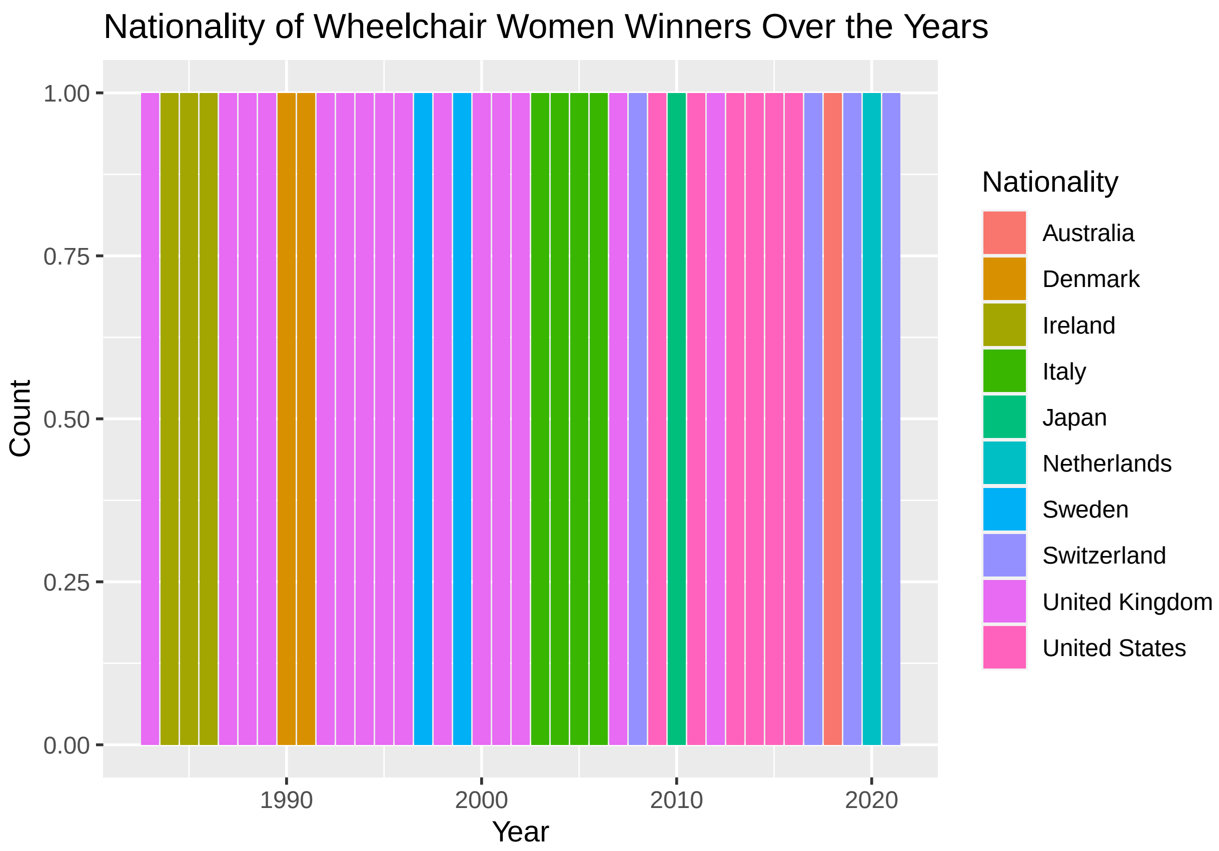
```r
ggplot(winners_data[winners_data$Category == "Women", ], aes(x = Year, fill = Nationality)) +
  geom_bar(position = "stack") +
  labs(title = "Nationality of Women Winners Over the Years",
       x = "Year",
       y = "Count",
       fill = "Nationality")
```



```r
ggplot(winners_data[winners_data$Category == "Wheelchair Men", ], aes(x = Year, fill = Nationality)) +
  geom_bar(position = "stack") +
  labs(title = "Nationality of Wheelchair Men Winners Over the Years",
       x = "Year",
       y = "Count",
       fill = "Nationality")
```



```
ggplot(winners_data[winners_data$Category == "Wheelchair Women", ], aes(x = Year, fill = Nationality)) +
  geom_bar(position = "stack") +
  labs(title = "Nationality of Wheelchair Women Winners Over the Years",
       x = "Year",
       y = "Count",
       fill = "Nationality")
```



Aspect 1: Winners' Nationalities

```
# Load required libraries
library(dplyr)


# Count the number of wins by nationality
nationality_counts <- winners_data %>%
  group_by(Category, Nationality) %>%
  summarise(Wins = n()) %>%
  arrange(desc(Wins))
```

```
## ‘summarise()‘ has grouped output by ’Category’. You can override using the
## ‘.groups‘ argument.
```

```
print("Winners' Nationalities:")
```

```
## [1] "Winners’ Nationalities:"
```

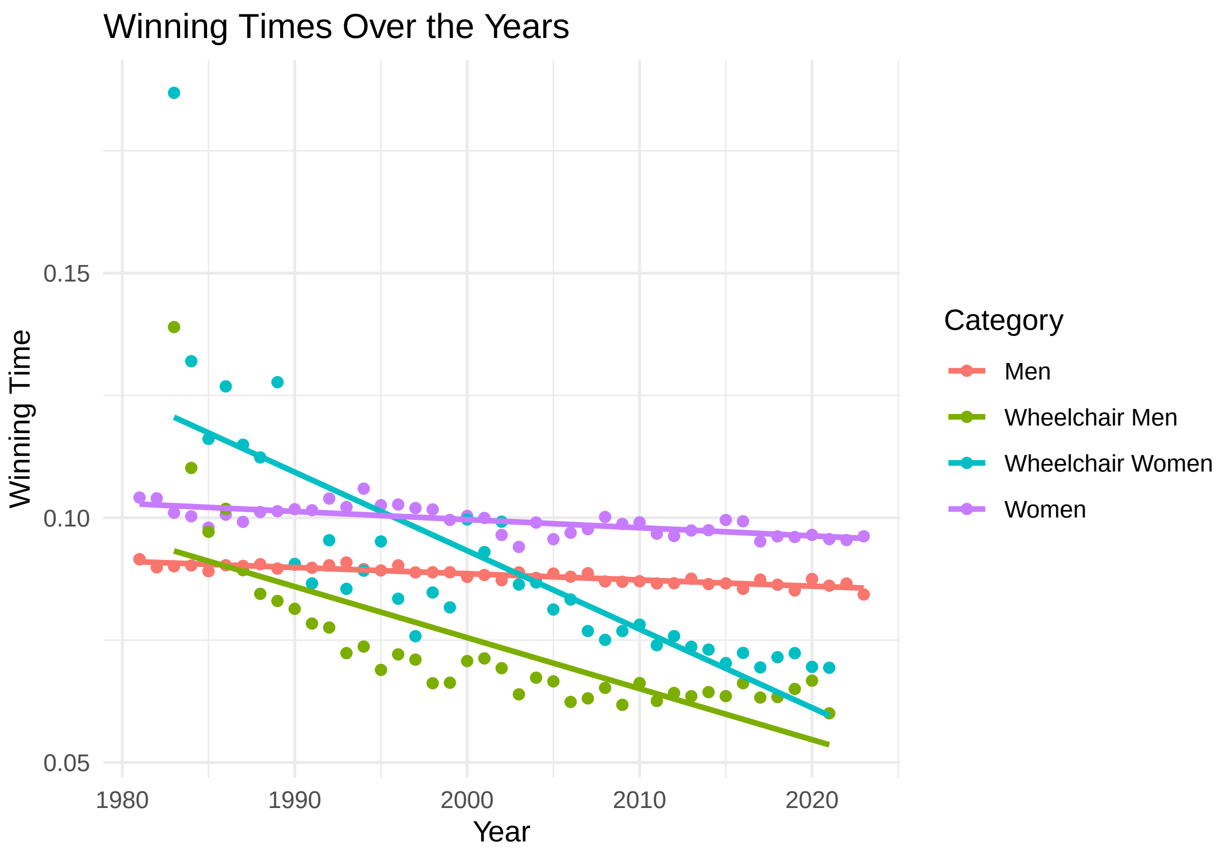
```
print(nationality_counts)
```

```
## # A tibble: 43 x 3
## # Groups:   Category [4]
##    Category         Nationality     Wins
##    <chr>            <chr>          <int>
##  1 Men              Kenya             17
##  2 Wheelchair Men   United Kingdom    16
##  3 Wheelchair Women United Kingdom    15
##  4 Women            Kenya             14
##  5 Women            United Kingdom     7
##  6 Men              United Kingdom     6
##  7 Wheelchair Men   Switzerland        6
##  8 Wheelchair Women United States      6
##  9 Women            Norway             6
## 10 Men              Ethiopia           5
## # i 33 more rows
```

Aspect 2: Winning Times

```
# Visualize winning times by category
library(ggplot2)

ggplot(winners_data, aes(x = Year, y = Time, color = Category)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Winning Times Over the Years",
       x = "Year",
       y = "Winning Time",
       color = "Category") +
  theme_minimal()
```

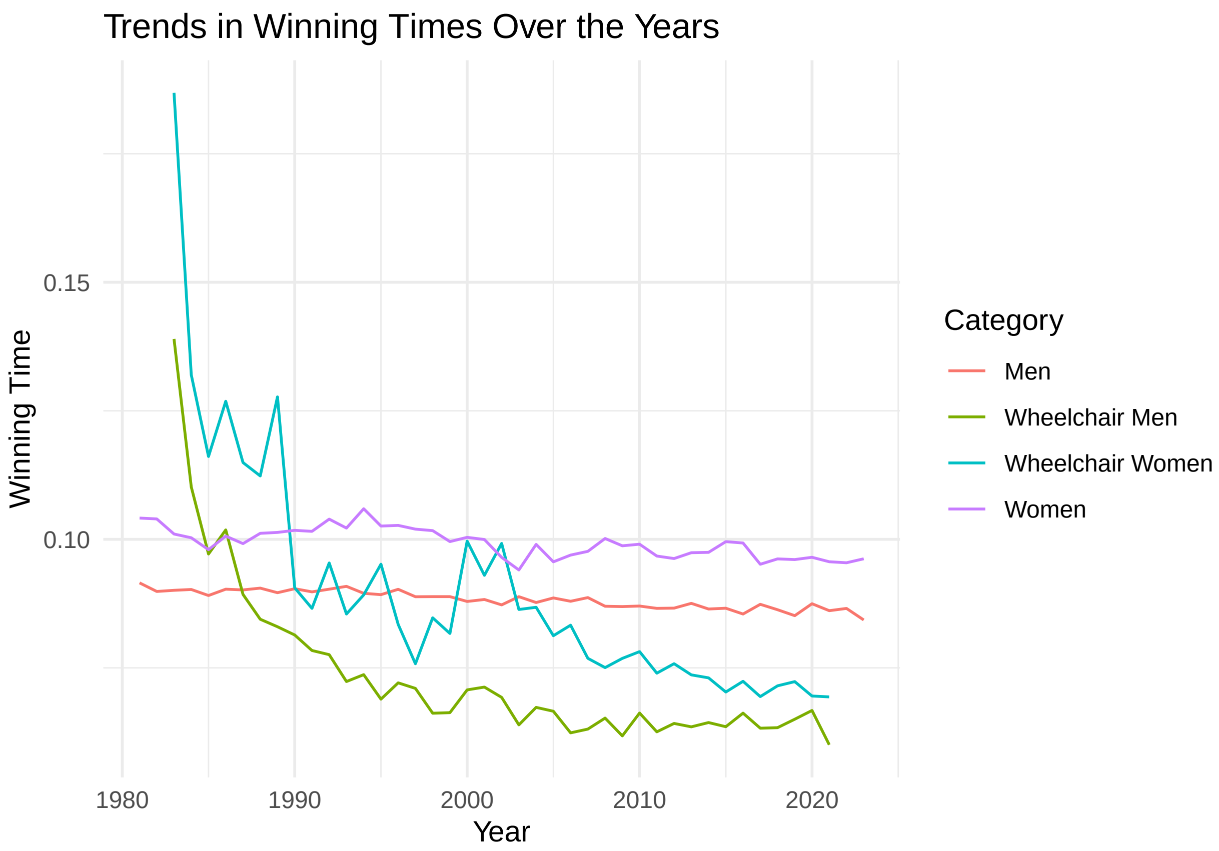
```
## Don’t know how to automatically pick scale for object of type <times>.
## Defaulting to continuous.
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



Aspect 3: Trends in Winning Times Over Years

```
# Visualize trends in winning times over the years
ggplot(winners_data, aes(x = Year, y = Time, color = Category)) +
  geom_line() +
  labs(title = "Trends in Winning Times Over the Years",
       x = "Year",
       y = "Winning Time",
       color = "Category") +
  theme_minimal()
```

```
## Don’t know how to automatically pick scale for object of type <times>.
## Defaulting to continuous.
```



Aspect 4: Winning Athletes

```
# Identify athletes with the most wins
top_athletes <- winners_data %>%
  group_by(Category, Athlete, Nationality) %>%
  summarise(Wins = n()) %>%
  arrange(desc(Wins))
```

```
## `summarise()` has grouped output by 'Category', 'Athlete'. You can override
## using the `.groups` argument.
```

```
print("Top Athletes with Most Wins:")
```

```
## [1] "Top Athletes with Most Wins:"
```

```
print(top_athletes)
```

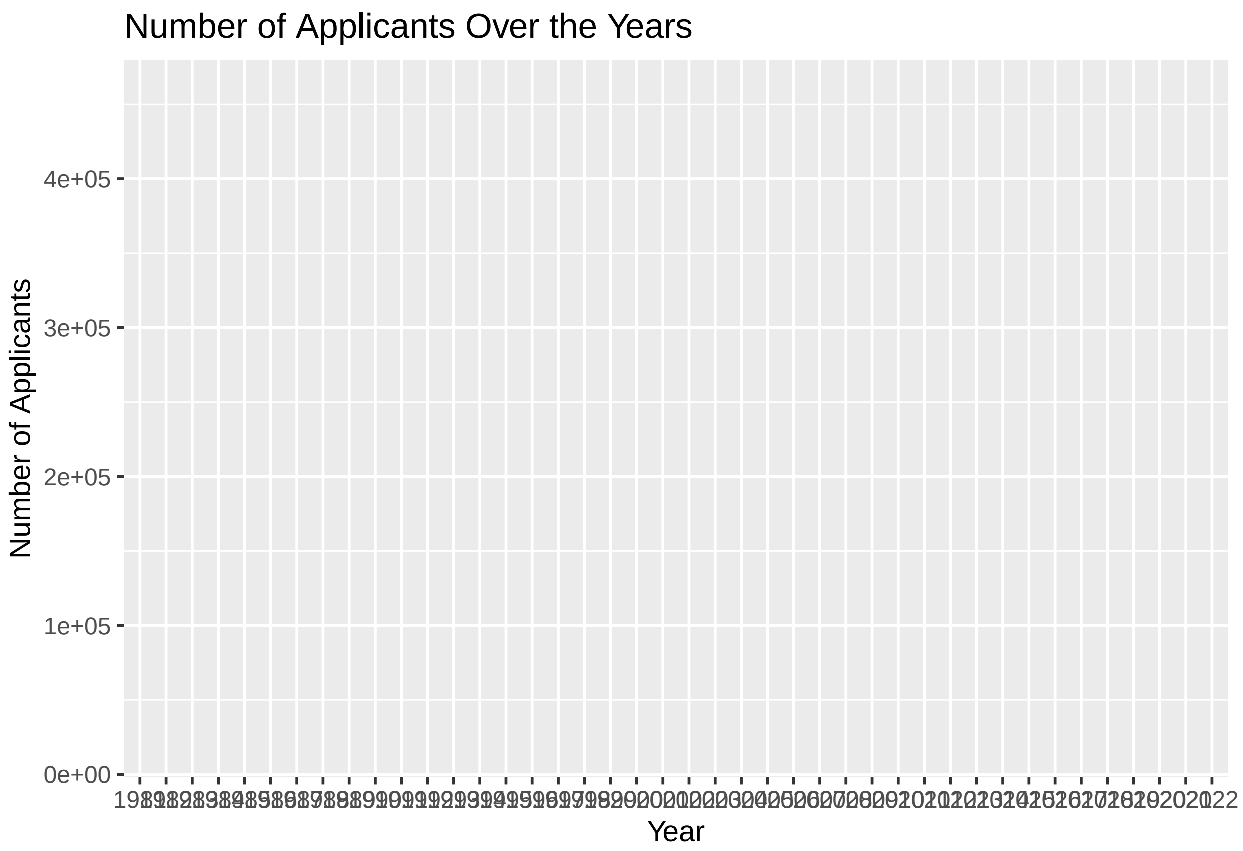
```
## # A tibble: 101 x 4
## # Groups:   Category, Athlete [101]
##    Category         Athlete              Nationality     Wins
##    <chr>            <chr>                <chr>          <int>
##  1 Wheelchair Men   David Weir           United Kingdom     8
##  2 Wheelchair Women Tanni Grey-Thompson  United Kingdom     6
##  3 Men              Eliud Kipchoge       Kenya              4
##  4 Wheelchair Men   David Holding        United Kingdom     4
##  5 Wheelchair Women Francesca Porcellato Italy              4
##  6 Wheelchair Women Tatyana McFadden     United States      4
##  7 Women            Ingrid Kristiansen   Norway             4
##  8 Men              António Pinto        Portugal           3
##  9 Men              Dionicio Cerón       Mexico             3
## 10 Men              Martin Lel           Kenya              3
## # i 91 more rows
```

Aspect 1: Winners' Nationalities

```r
library(ggplot2)

# Convert the 'Year' column to a factor to maintain the order
marathon_data$Year <- as.factor(marathon_data$Year)

# Plotting the trend of applicants over the years
ggplot(marathon_data, aes(x = Year, y = Applicants)) +
  geom_line() +
  labs(title = "Number of Applicants Over the Years",
       x = "Year",
       y = "Number of Applicants")
```

```
## Warning: Removed 2 rows containing missing values (`geom_line()`).
```

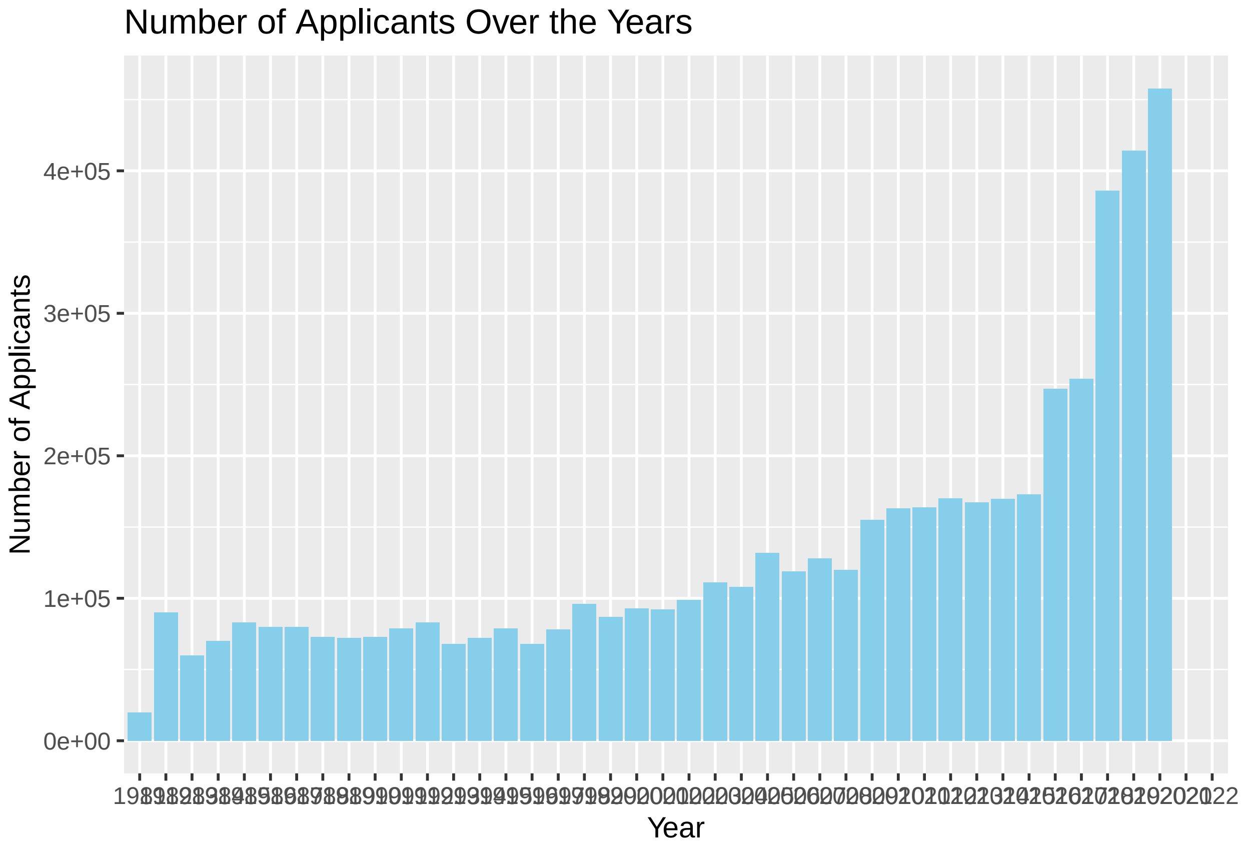
```
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```



Bar plot

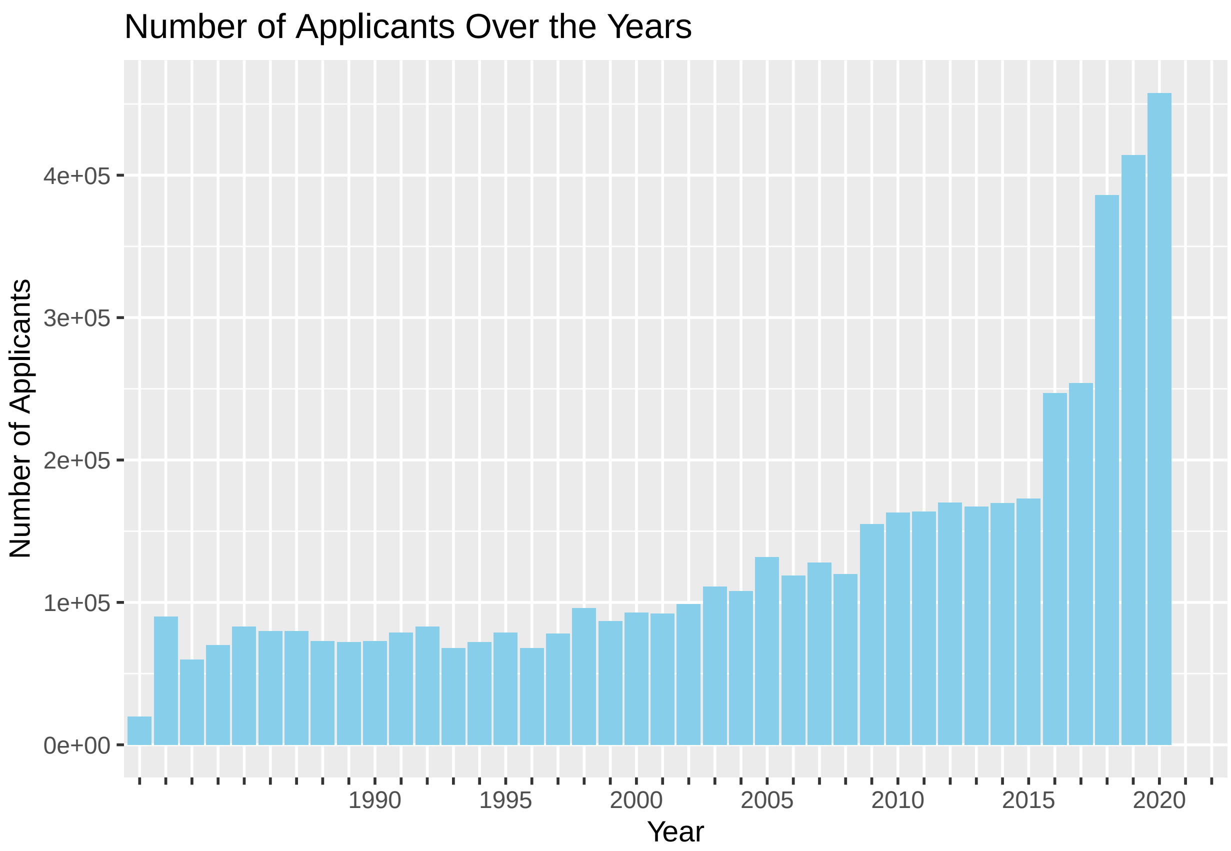
```
# Bar plot of the number of applicants over the years
ggplot(marathon_data, aes(x = Year, y = Applicants)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Number of Applicants Over the Years",
       x = "Year",
       y = "Number of Applicants")
```

```
## Warning: Removed 2 rows containing missing values (‘position_stack()‘).
```



```
# Define the subset of years for better readability
subset_years <- c(1990, 1995, 2000, 2005, 2010, 2015, 2020)

# Bar plot of the number of applicants over the years with adjusted axis labels
ggplot(marathon_data, aes(x = factor(Year), y = Applicants)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Number of Applicants Over the Years",
       x = "Year",
       y = "Number of Applicants") +
  scale_x_discrete(labels = function(x) ifelse(as.numeric(x) %in% subset_years, x, ""))
```

```
## Warning: Removed 2 rows containing missing values (‘position_stack()‘).
```

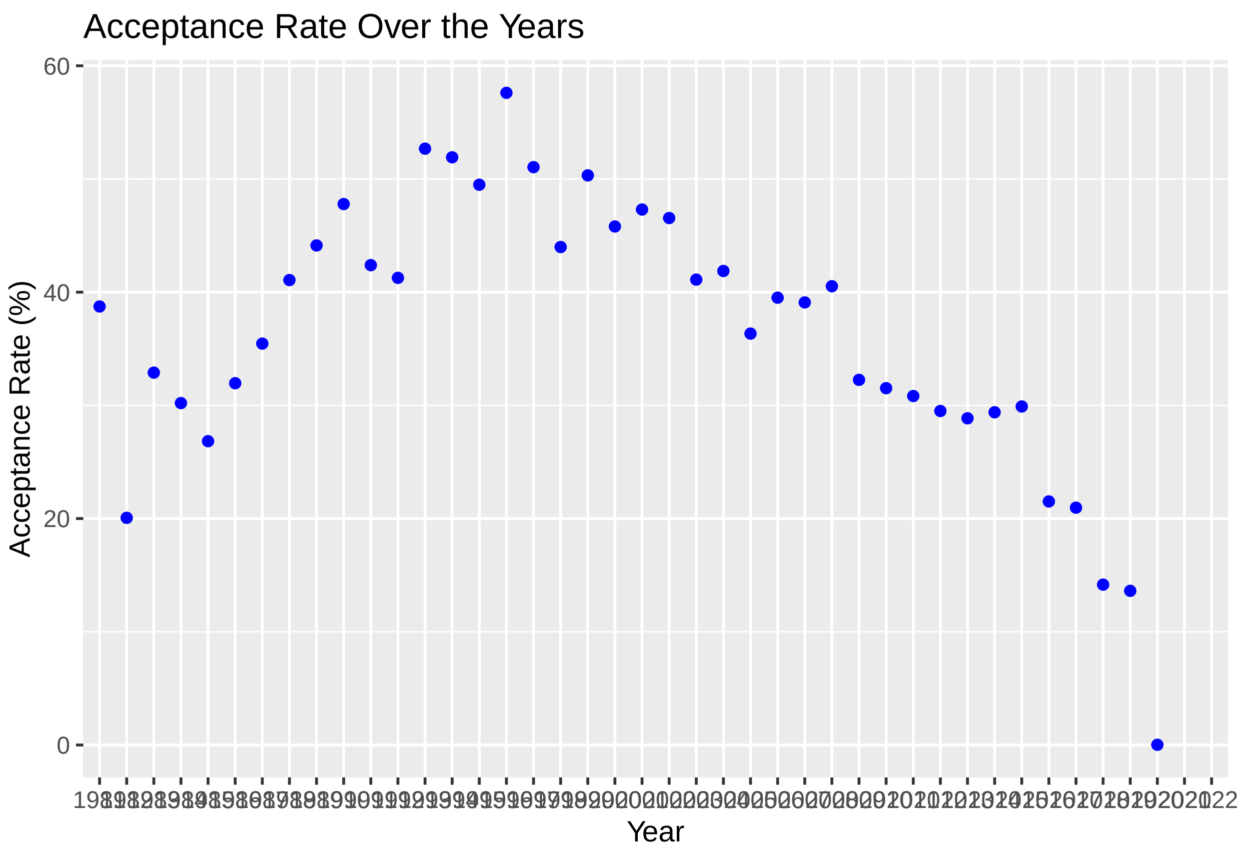


```r
# Calculate acceptance rate
marathon_data$Acceptance_Rate <- (marathon_data$Accepted / marathon_data$Applicants) * 100

# Line plot of acceptance rate over the years
ggplot(marathon_data, aes(x = Year, y = Acceptance_Rate)) +
  geom_line(color = "blue") +
  geom_point(color = "blue") +
  labs(title = "Acceptance Rate Over the Years",
       x = "Year",
       y = "Acceptance Rate (%)")
```

```
## Warning: Removed 2 rows containing missing values (`geom_line()`).
```

```
## `geom_line()`: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```

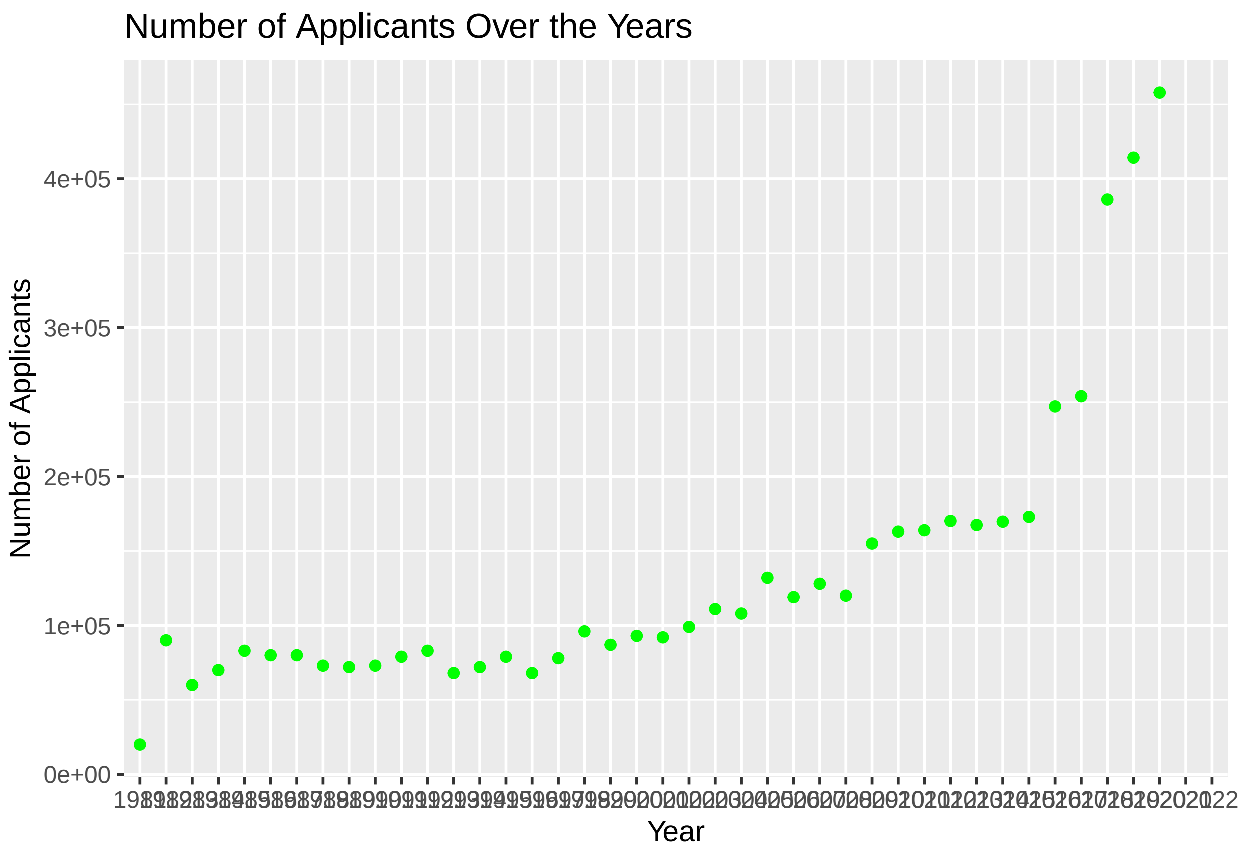


1)Number of Applicants Over the Years:

Visualized using a line plot to show the trend. Increasing trend indicates growing interest in the marathon.

```
ggplot(marathon_data, aes(x = Year, y = Applicants)) +
  geom_line(color = "green") +
  geom_point(color = "green") +
  labs(title = "Number of Applicants Over the Years",
       x = "Year",
       y = "Number of Applicants")
```

```
## Warning: Removed 2 rows containing missing values (‘geom_line()‘).

## ‘geom_line()‘: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?

## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



2)Number of Finishers Over the Years:

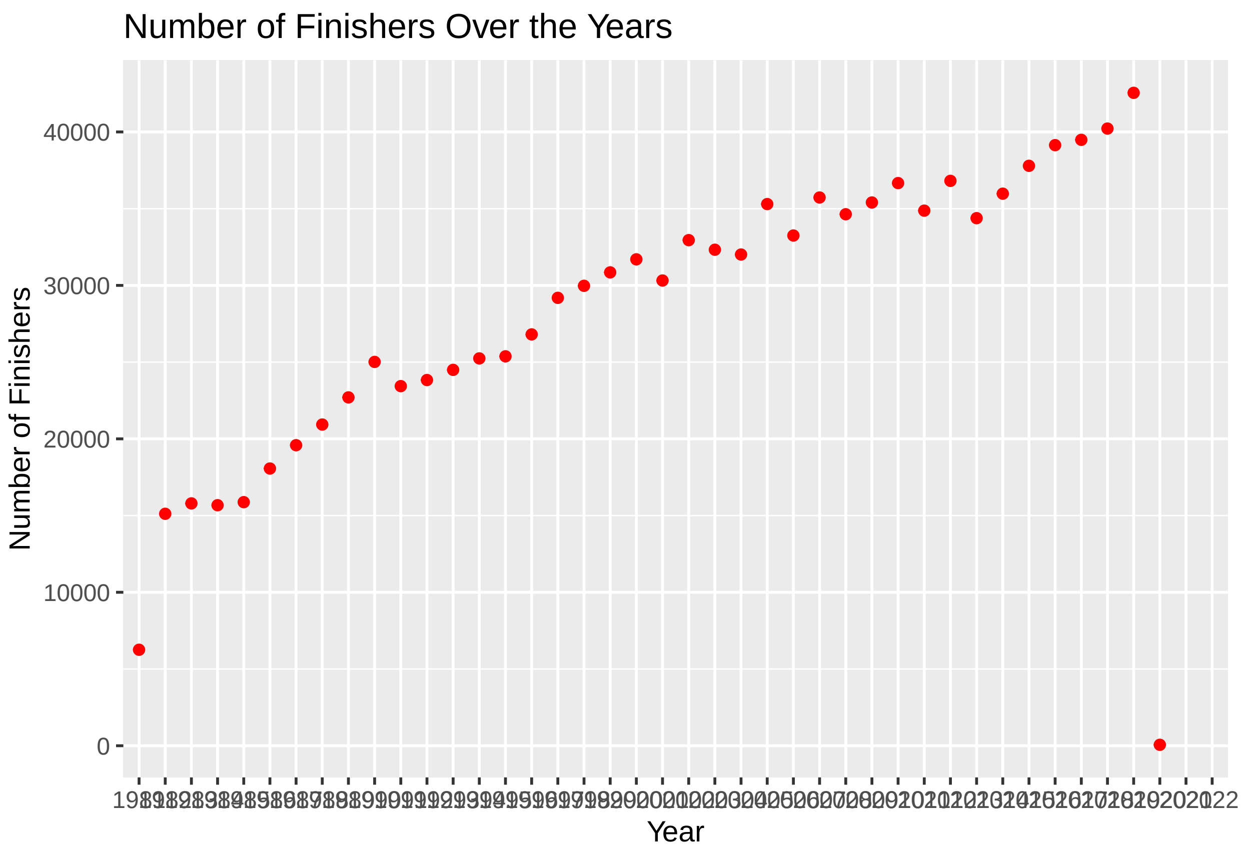
Visualized using a line plot. Indicates the growth in the number of participants who successfully completed the marathon.

```
ggplot(marathon_data, aes(x = Year, y = Finishers)) +
  geom_line(color = "red") +
  geom_point(color = "red") +
  labs(title = "Number of Finishers Over the Years",
       x = "Year",
       y = "Number of Finishers")
```

```
## Warning: Removed 2 rows containing missing values (‘geom_line()‘).
```

```
## ‘geom_line()‘: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```

```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



3)Acceptance Rate Over the Years:

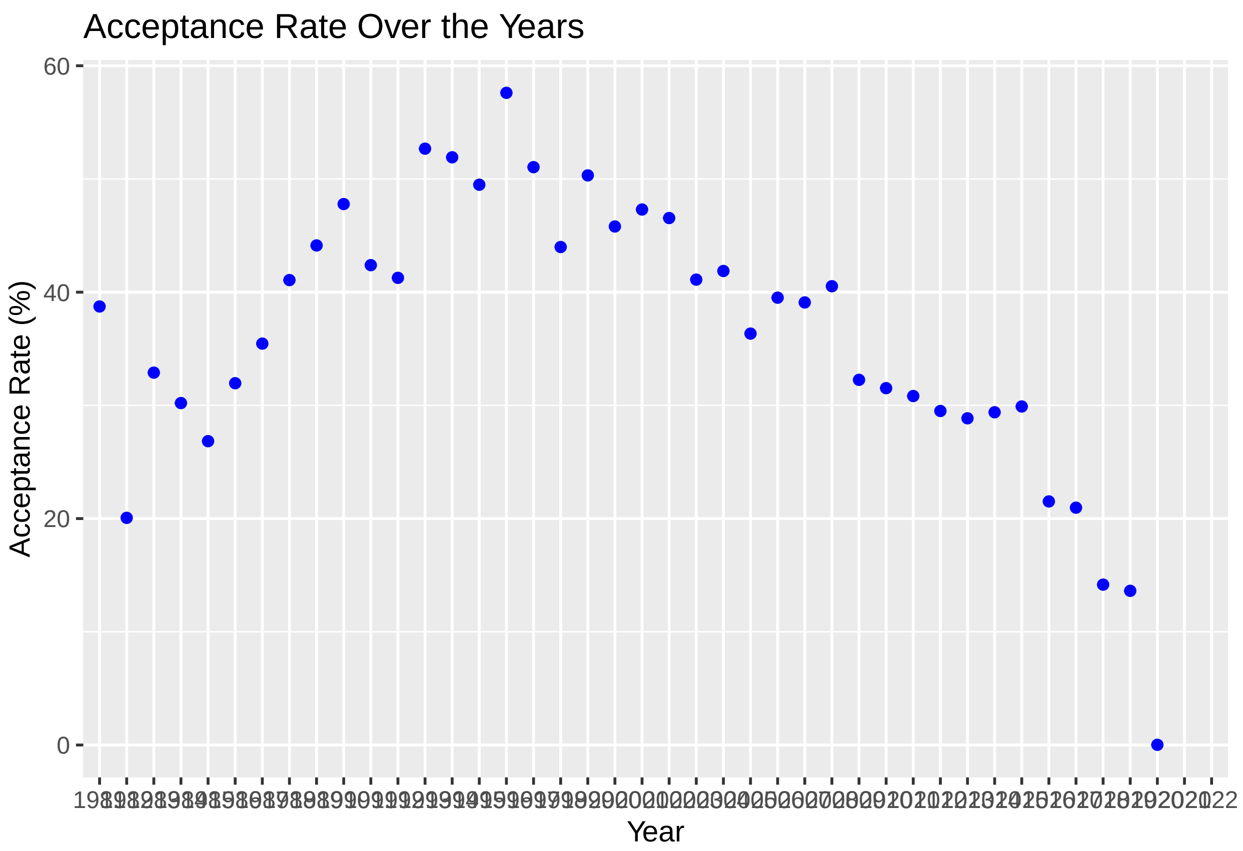
Calculated and visualized using a line plot. Helps understand the competitiveness of the marathon.

```r
marathon_data$Acceptance_Rate <- (marathon_data$Accepted / marathon_data$Applicants) * 100

ggplot(marathon_data, aes(x = Year, y = Acceptance_Rate)) +
  geom_line(color = "blue") +
  geom_point(color = "blue") +
  labs(title = "Acceptance Rate Over the Years",
       x = "Year",
       y = "Acceptance Rate (%)")
```

```
## Warning: Removed 2 rows containing missing values (‘geom_line()‘).
```

```
## ‘geom_line()‘: Each group consists of only one observation.
## i Do you need to adjust the group aesthetic?
```

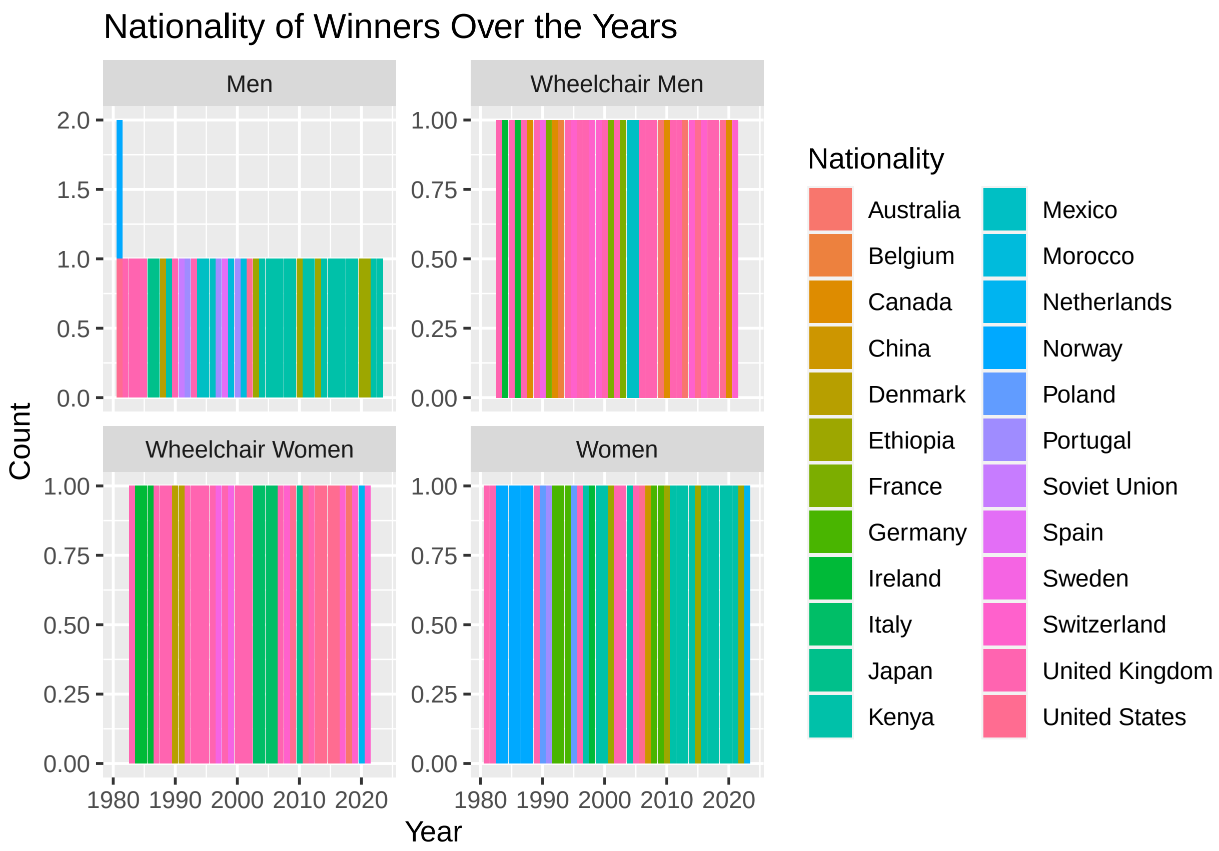
```
## Warning: Removed 2 rows containing missing values (‘geom_point()‘).
```



4)Nationality of Winners Over the Years (by Category):

Visualized using a stacked bar plot with facets for each category. Provides insights into the diversity of winners by nationality.

```
ggplot(winners_data, aes(x = Year, y = ..count.., fill = Nationality)) +
  geom_bar(position = "stack") +
  labs(title = "Nationality of Winners Over the Years",
       x = "Year",
       y = "Count",
       fill = "Nationality") +
  facet_wrap(~Category, scales = "free_y")
```

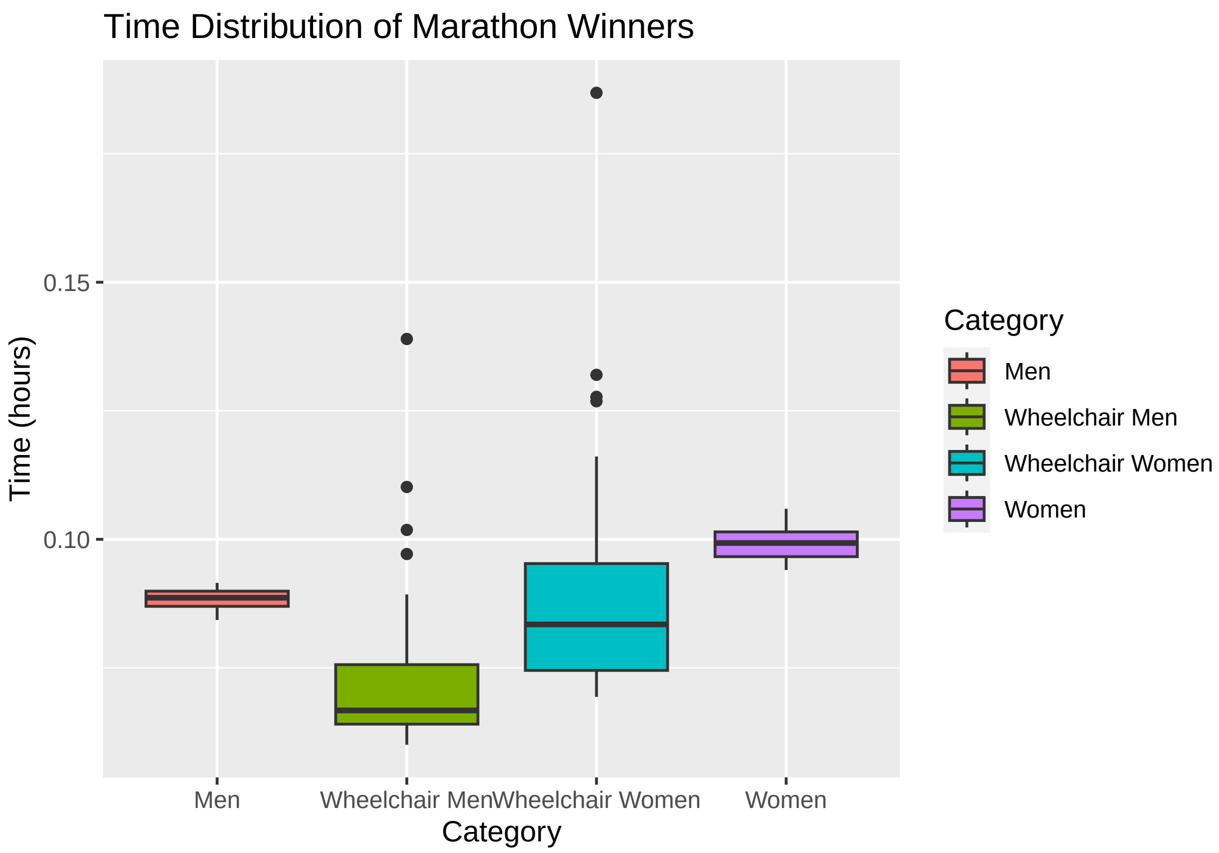


5)Time Distribution of Marathon Winners (by Category):

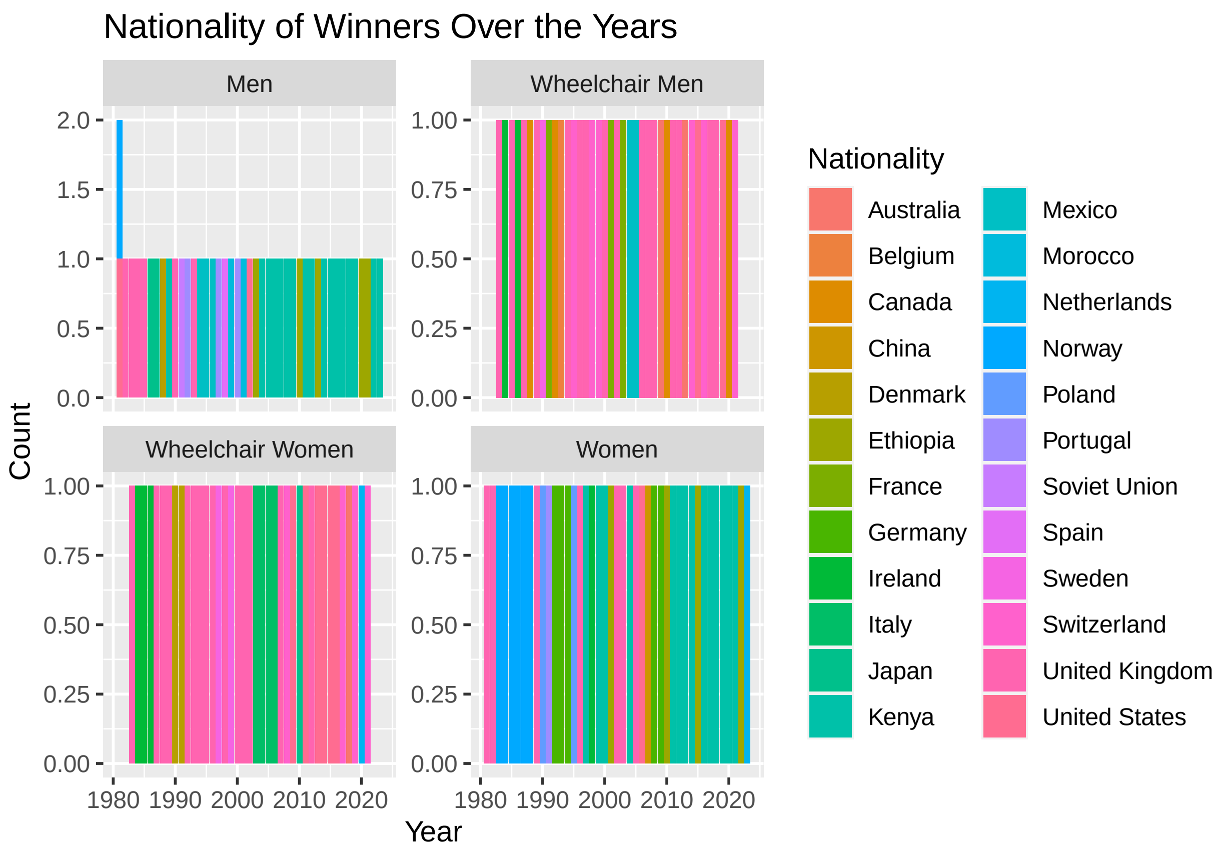
Box plot to show the distribution of winning times. Helps identify trends in performance over the years.

```
ggplot(winners_data, aes(x = Category, y = Time, fill = Category)) +
  geom_boxplot() +
  labs(title = "Time Distribution of Marathon Winners",
       x = "Category",
       y = "Time (hours)",
       fill = "Category")
```

```
## Don't know how to automatically pick scale for object of type <times>.
## Defaulting to continuous.
```



```r
ggplot(winners_data, aes(x = Year, y = ..count.., fill = Nationality)) +
  geom_bar(position = "stack") +
  labs(title = "Nationality of Winners Over the Years",
       x = "Year",
       y = "Count",
       fill = "Nationality") +
  facet_wrap(~Category, scales = "free_y")
```


Nationality of Winners Over the Years
Men
Wheelchair Men
Wheelchair Women
Women
Count
Year
2.0
1.5
1.0
0.5
0.0
1.00
0.75
0.50
0.25
0.00
1980
1990
2000
2010
2020
Nationality
Australia
Belgium
Canada
China
Denmark
Ethiopia
France
Germany
Ireland
Italy
Japan
Kenya
Mexico
Morocco
Netherlands
Norway
Poland
Portugal
Soviet Union
Spain
Sweden
Switzerland
United Kingdom
United States