# Mens and Womens Boston Marathon Winners

Ida Mazinger

06.06.2023

## Geschwindigkeiten

Alle Geschwindigkeitsangaben sind in Minuten pro Kilometer.

```
# Geschwindigkeiten von Männern und Frauen
all_times <- format(as.POSIXct(strptime(all_data$Time, format="%H:%M:%S")),
                    format = "%H:%M:%S")
all_times_hms <- as_hms(all_times) # Umwandlung in Stunden-Minuten-Sekunden
all_times_in_seconds <- lubridate::seconds(all_times_hms)
all_times_in_minutes <- as.numeric(all_times_in_seconds / 60)
all_speeds <- all_times_in_minutes / all_data$Distance..KM.
filtered_all_speeds <- c(all_speeds)[!(is.null(all_speeds) | is.na(all_speeds) | all_speeds == "")]

# Geschwindigkeiten von Frauen
womens_times <- format(as.POSIXct(strptime(womens_data$Time, format = "%H:%M:%S")),
                    format = "%H:%M:%S")
womens_times_hms <- as_hms(womens_times) # Umwandlung in Stunden-Minuten-Sekunden
womens_times_in_seconds <- lubridate::seconds(womens_times_hms)
womens_times_in_minutes <- as.numeric(womens_times_in_seconds / 60)
womens_speeds <- womens_times_in_minutes / womens_data$Distance..KM.

# Geschwindigkeiten von Männern
mens_times <- format(as.POSIXct(strptime(mens_data$Time, format="%H:%M:%S")),
                    format = "%H:%M:%S")
mens_times_hms <- as_hms(mens_times) # Umwandlung in Stunden-Minuten-Sekunden
mens_times_in_seconds <- lubridate::seconds(mens_times_hms)
mens_times_in_minutes <- as.numeric(mens_times_in_seconds / 60)
mens_speeds <- mens_times_in_minutes / mens_data$Distance..KM.
```

### Durchschnitt:

```
mean_speeds <- mean(all_speeds, na.rm = TRUE)
mean_womens_speeds <- mean(womens_speeds, na.rm = TRUE)
mean_mens_speeds <- mean(mens_speeds, na.rm = TRUE)
diff_means <- mean_mens_speeds - mean_womens_speeds
```

Die Durchschnittsgeschwindigkeit beträgt etwa 3.49 min/km. Die Durchschnittsgeschwindigkeit der Frauen beträgt etwa 3.68 min/km. Die Durchschnittsgeschwindigkeit der Männer beträgt etwa 3.41 min/km. Männer sind im Schnitt um -0.27 min/km schneller.

**Median:**

```
median_speeds <- median(all_speeds, na.rm = TRUE)
median_womens_speeds <- median(womens_speeds, na.rm = TRUE)
median_mens_speeds <- median(mens_speeds, na.rm = TRUE)
diff_median <- abs(mean_womens_speeds -  mean_mens_speeds)
```

Der Median beträgt etwa 3.44 min/km. Der Median der Frauen beträgt etwa 3.46 min/km. Der Median der Männer beträgt etwa 3.44 min/km. Der Unterschied zwischen den Medianen beträgt 0.27 min/km.

**Minimum und Maximum:**

```
min_speed <- min(all_speeds, na.rm = TRUE)
min_womens_speed <- min(womens_speeds, na.rm = TRUE)
min_mens_speed <- min(mens_speeds, na.rm = TRUE)
# schnellste Geschwindigkeit entspricht der der Männer:
min_speed == min_mens_speed
```

```
## [1] TRUE
```

```
max_speed <- max(all_speeds, na.rm = TRUE)
max_womens_speed <- max(womens_speeds, na.rm = TRUE)
max_mens_speed <- max(mens_speeds, na.rm = TRUE)
# langsamste Geschwindigkeit entspricht der der Frauen:
max_speed == max_womens_speed
```

```
## [1] TRUE
```

Die schnellste Geschwindigkeit der Frauen beträgt etwa 3.32 min/km. Die schnellste Geschwindigkeit der Männer beträgt etwa 2.92 min/km und ist auch die generelle minimale Geschwindigkeit.

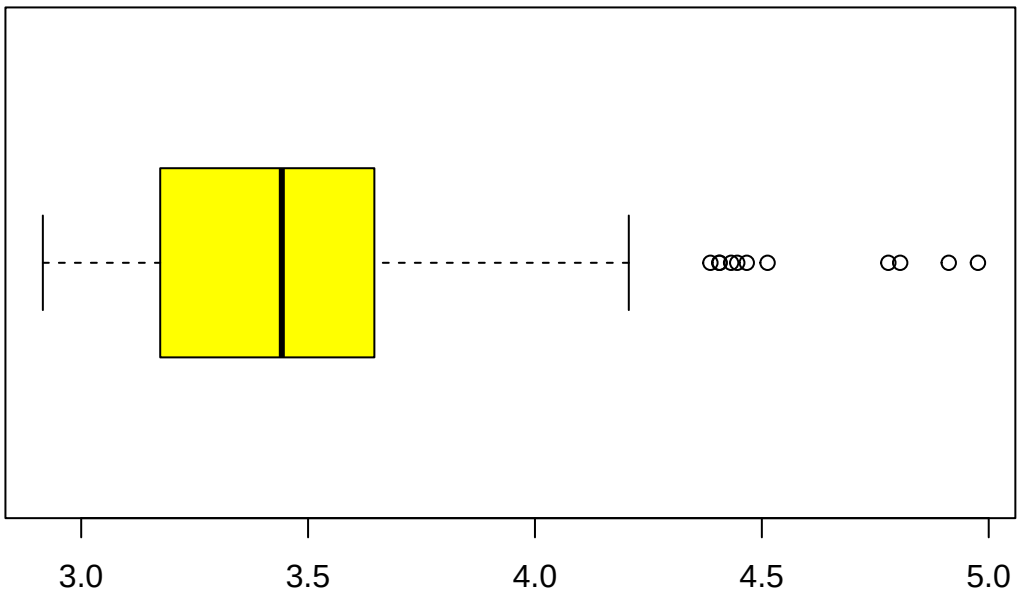
Die langsamste Geschwindigkeit der Frauen beträgt etwa 4.98 min/km und ist auch die generelle maximale Geschwindigkeit. Die langsamste Geschwindigkeit der Männer beträgt etwa 4.45 min/km.

**Boxplot der Geschwindigkeiten:**

```
quartiles_all_speeds <- quantile(filtered_all_speeds, probs = c(0, 0.25, 0.5, 0.75, 1))
summary_all_speeds <- summary(all_speeds)
summary_all_speeds
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   2.915   3.176   3.442   3.494   3.644   4.976       3
```

```
boxplot(all_speeds, type=1, col=c('yellow'), horizontal = TRUE)
```

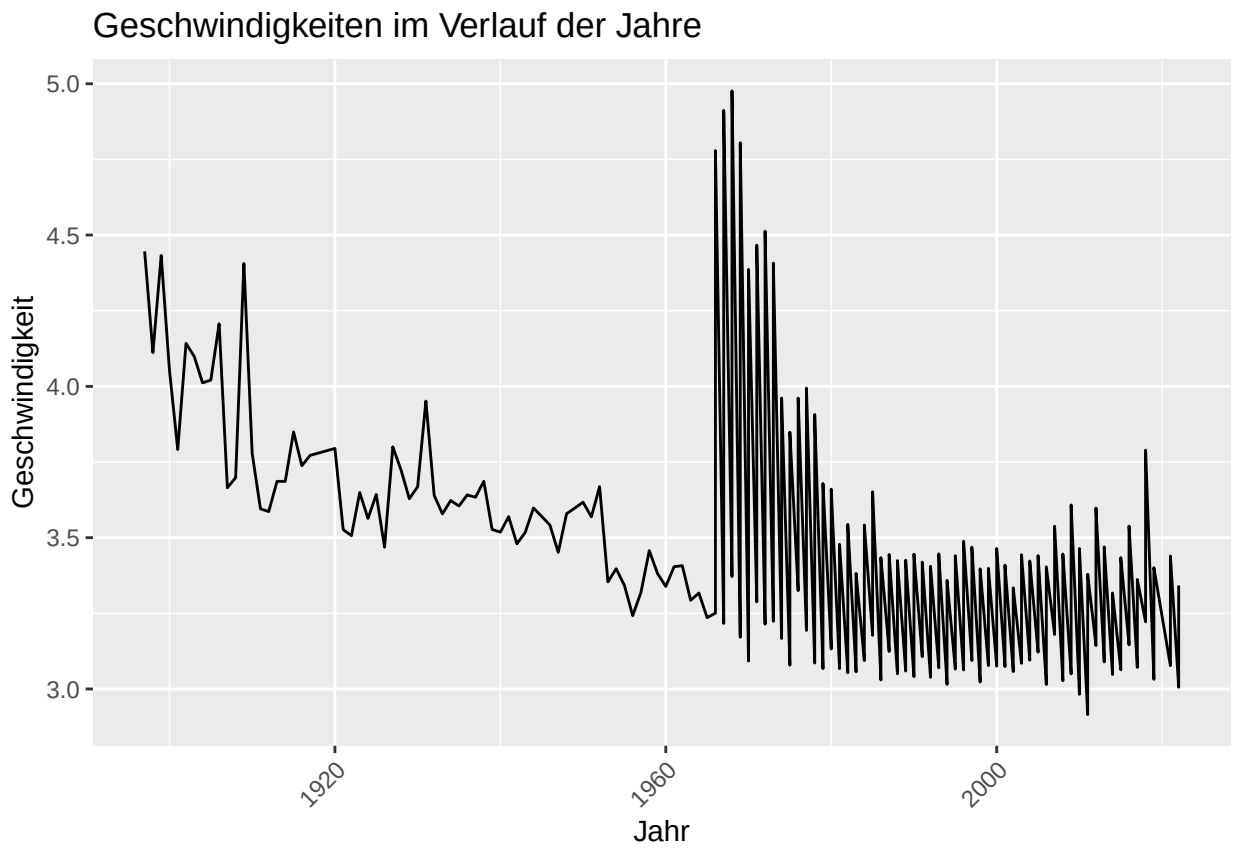



Der Boxplot zeigt, dass 50 % der Abschnittsgeschwindigkeiten zwischen ca. 3,18 min/km und 3,64 min/km liegen. Der Boxplot ist rechtsschief, das bedeutet, dass die meisten Gewinner eine recht schnell gelaufen sind (bis zu etwa 4,2 min/km) und nur ein paar eher langsam. Da die Maximalgeschwindigkeit der Männer bei etwa 4.45 liegt, können wir davon ausgehen, dass die meisten Ausreißer vom Marathond der Frauen stammen.

### Jahreszahl im Vergleich zu Geschwindigkeit

```
# barplot(all_speeds, col = rainbow(250), main = "Geschwindigkeiten im Verlauf der Jahre")
all_speeds_df <- data.frame(Jahr = all_data$Year,
                            Geschwindigkeit = all_speeds)
# todo wenn mehrere Werte pro Jahr: Durchschnitt
ggplot(all_speeds_df, aes(x = Jahr, y = Geschwindigkeit)) +
  geom_line() +
  xlab("Jahr") +
  ylab("Geschwindigkeit") +
  ggtitle("Geschwindigkeiten im Verlauf der Jahre") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 3 rows containing missing values (‘geom_line()‘).
```

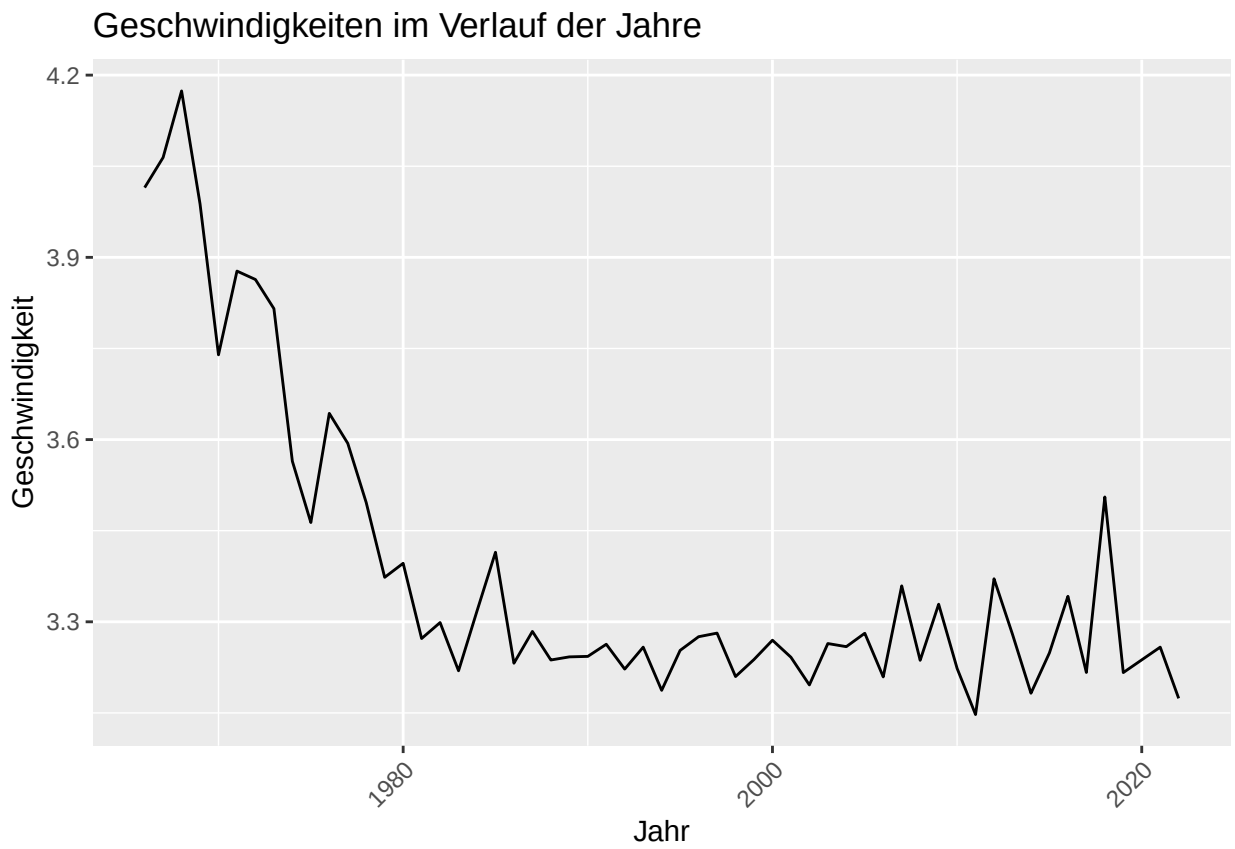



Das Liniendiagramm lässt klar erkennen, dass die Geschwindigkeit im Laufe der ersten Jahre insgesamt stark abgenommen hat. Ab dem Jahr 1966 gibt es auch Daten über Frauenläufe, daher ist der Graf ab hier kaum leserlich. Im folgenden nehmen wir daher alle Daten ab 1966, berechnen jeweils den Durchschnitt und stellen diesen dar:

```
mens_data_since_1966 <- subset(mens_data, Year >= 1966)
filtered_womens_data <- womens_data[!(is.na(womens_data$Year) | is.nan(womens_data$Year)), ]
merged_data <- merge(mens_data_since_1966, filtered_womens_data, by = "Year")
merged_data$Speed.x <- as.numeric(lubridate::seconds(as_hms(format(as.POSIXct(strptime(merged_data$Time
merged_data$Speed.y <- as.numeric(lubridate::seconds(as_hms(format(as.POSIXct(strptime(merged_data$Time
mean_speed_since_1966 <- rowMeans(merged_data[, c("Speed.x", "Speed.y")], na.rm = TRUE)

mean_speeds_since_1966_df <- data.frame(Jahr = merged_data$Year,
                            Geschwindigkeit = mean_speed_since_1966)

ggplot(mean_speeds_since_1966_df, aes(x = Jahr, y = Geschwindigkeit)) +
  geom_line() +
  xlab("Jahr") +
  ylab("Geschwindigkeit") +
  ggtitle("Geschwindigkeiten im Verlauf der Jahre") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




Im neuen Liniendiagramm sieht man die Durchschnittsabschnittsgeschwindigkeit von Männern und Frauen seit 1966.

### Beziehung zwischen Zeit und Entfernung

```
# todo
```

### Anzahl Gewinner pro Land

```
frequency <- table(all_data$Country, exclude = c(NULL, ""))
df <- data.frame(Land = names(frequency), Häufigkeit = as.vector(unname(frequency)))
df <- arrange(df, desc(frequency))
df
```

```
##              Land Häufigkeit
## 1   United States         59
## 2           Kenya         38
## 3          Canada         17
## 4        Ethiopia         14
## 5           Japan          9
## 6         Finland          7
## 7         Germany          6
## 8          Russia          4
```

```
## 9     New Zealand          3
## 10       Portugal          3
## 11    South Korea          3
## 12 United Kingdom          3
## 13        Belgium          2
## 14         Greece          2
## 15         Norway          2
## 16      Australia          1
## 17       Colombia          1
## 18      Guatemala          1
## 19        Ireland          1
## 20          Italy          1
## 21         Poland          1
## 22         Sweden          1
## 23     Yugoslavia          1
```

```
#pie(df_filtered$Frequency.Freq, labels = df_filtered$Country, main = "Wins by Country")
```

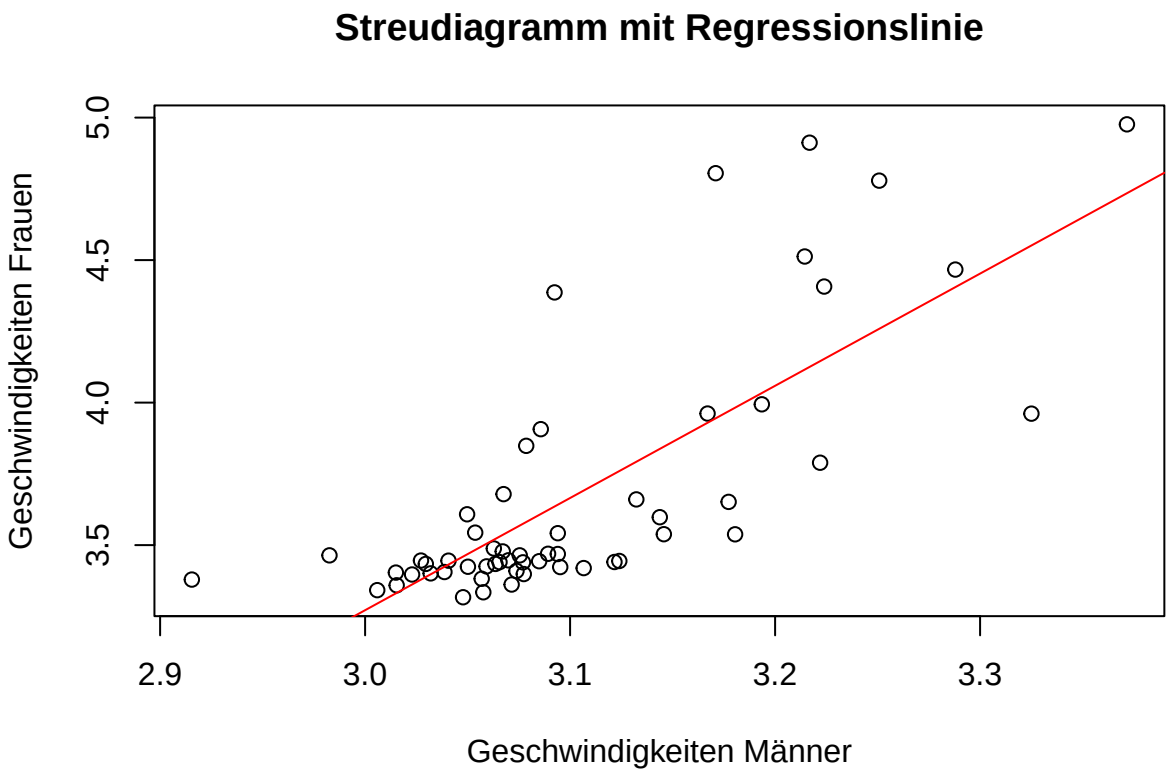
In der Tabelle wird ersichtlich, dass die meisten Gewinner aus den Vereinigten Staaten, Kenya und Kanada stammen.

### Streudiagramm?

Linie Durchschnittsgeschwindigkeit Punkte von Geschwindigkeiten v. Frauen und Männern (unterschiedliche Farben) ab Jahr von ca. 1960 wenn Frauen auch dabei sind

```
lm_model <- lm(Speed.y ~ Speed.x, data = merged_data)

predicted_values <- predict(lm_model)
plot(
  merged_data$Speed.x,
  merged_data$Speed.y,
  main = "Streudiagramm mit Regressionslinie",
  xlab = "Geschwindigkeiten Männer",
  ylab = "Geschwindigkeiten Frauen")
abline(lm_model, col = "red")
```




```r
covariance <- cov(merged_data$Speed.x, merged_data$Speed.y)
correlation <- cor(merged_data$Speed.x, merged_data$Speed.y)
```

Im Streudiagramm sieht man, wie sich die Geschwindigkeiten von Frauen und Männer zueinander verhalten. Interessant ist, dass die Leistungen sich leicht ähnlich verhalten. In Jahren, in denen Männer eher schwach abgeschnitten haben, was das auch bei den Frauen so und umgekehrt (siehe Regressionslinie). Die Kovarianz beträgt 0.0287. Dies bedeutetet einen positiven Zusammenhang zwischen den Geschwindigkeiten. Der Korrelationskoeffizient beträgt 0.7606 und weist darauf hin, dass die Daten recht stark positiv korellieren.

### Standardabweichungen

```r
# todo
```