# Homework 1

Anishka Chauhan

2022-06-29

1)

a)

```
x = c(16, 30, -6, 8, 18, 2, -22, 15, -68, 86)
y = c(87, -7, 50, -3, -112, 45, 1, -4, 9, 22)
```

b)

```
mean(x)
```

```
## [1] 7.9
```

```
mean(y)
```

```
## [1] 8.8
```

c)

```
xd = x - mean(x)
yd = y - mean(y)
```

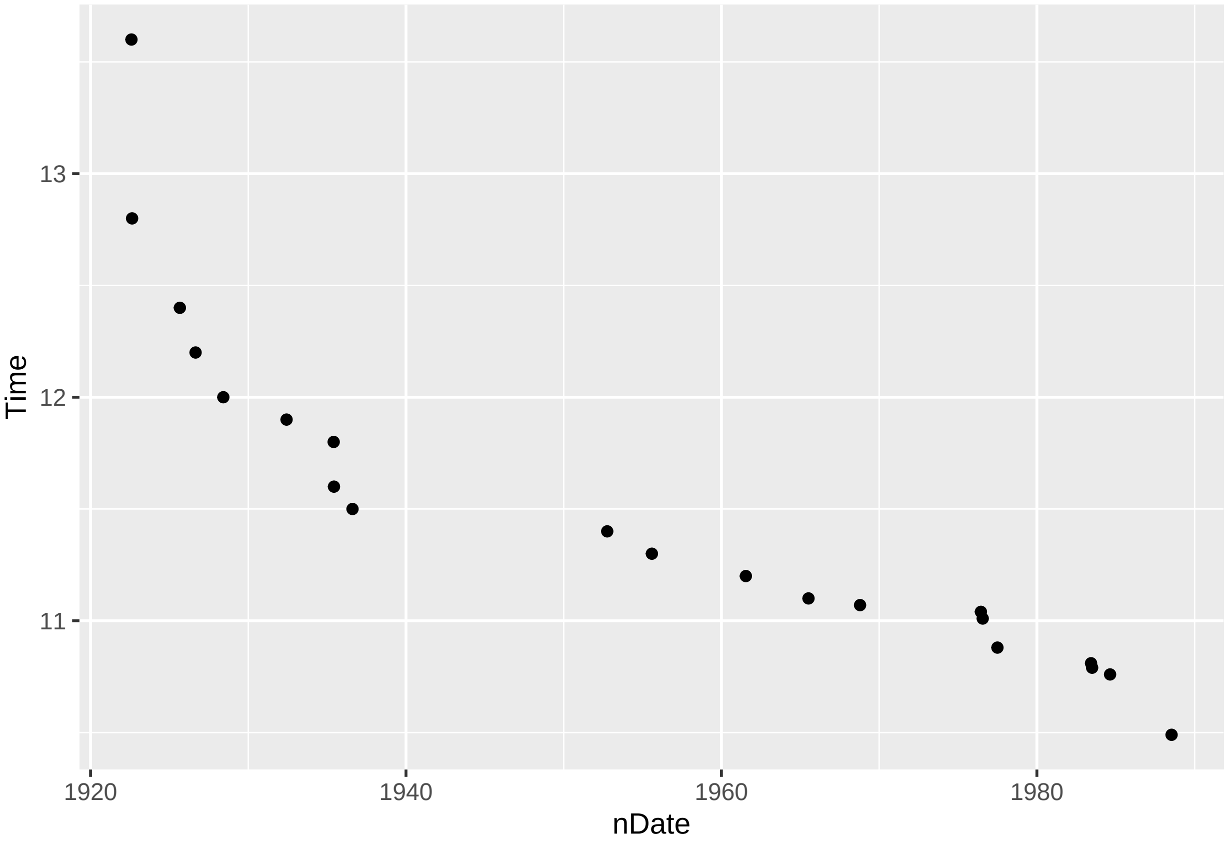
2) Working reproducibily helps with rooting out sources of error that could potentially undermine a computation-based science. It also facilitates better teamwork as it requires communication about the development process of the study, and allows other researchers to use the methods developed for their own studies instead of going through the expense of creating their own for the same tasks.

3)

```
require('ggplot2')
```

```
## Loading required package: ggplot2
```

```
men100 <- read.csv("./data/100men.csv",sep='\t')
women100 <- read.csv("./data/100women.csv",sep='\t')
women100$nDate <- as.Date(as.character(women100$Date),format='%b %d, %Y')
ggplot(women100,aes(x=nDate,y=Time)) + geom_point()
```



a)

```
# women100$Nationality = gsub(" ", "", women100$Nationality)
library(stringr)
women100$Nationality = str_trim(women100$Nationality)
women100$Nationality = gsub("West Germany", "Germany", women100$Nationality)
women100$Nationality = gsub("East Germany", "Germany", women100$Nationality)
women100$Nationality = gsub("United States", "USA", women100$Nationality)

men100$Nation = gsub("West Germany", "Germany", men100$Nation)

womenNationality = women100$Nationality

menNationality = men100$Nation
```

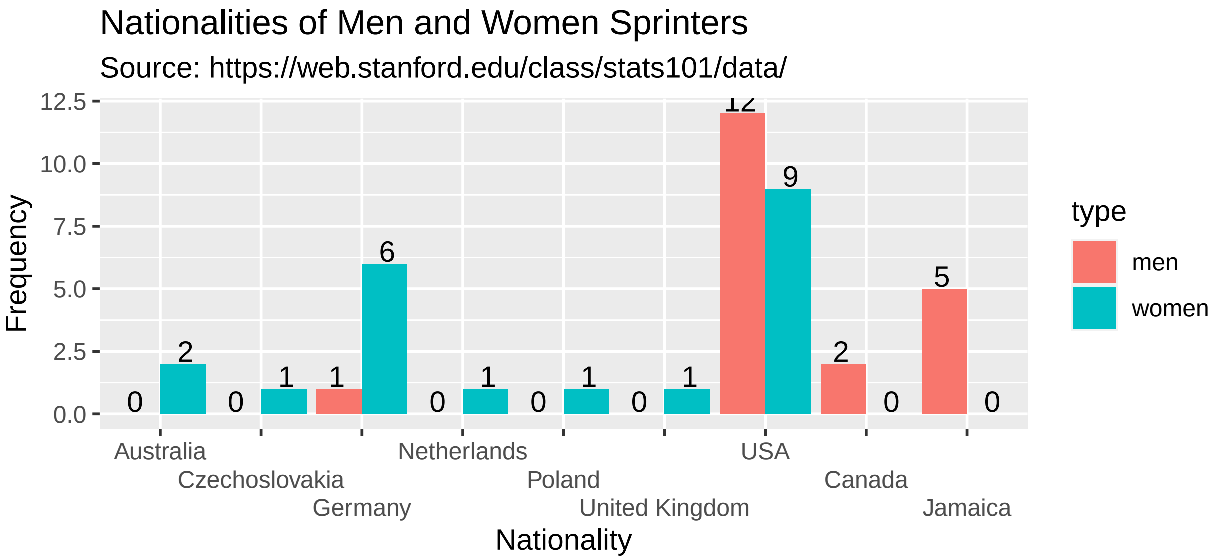
b)

```
w = table(womenNationality)
w = as.data.frame(w)
colnames(w)[1] = "nationality"

m = table(menNationality)
m = as.data.frame(m)
colnames(m)[1] = "nationality"

b_nation= merge(w, m, by = "nationality", all = T)
b_nation[is.na(b_nation)] = 0
```

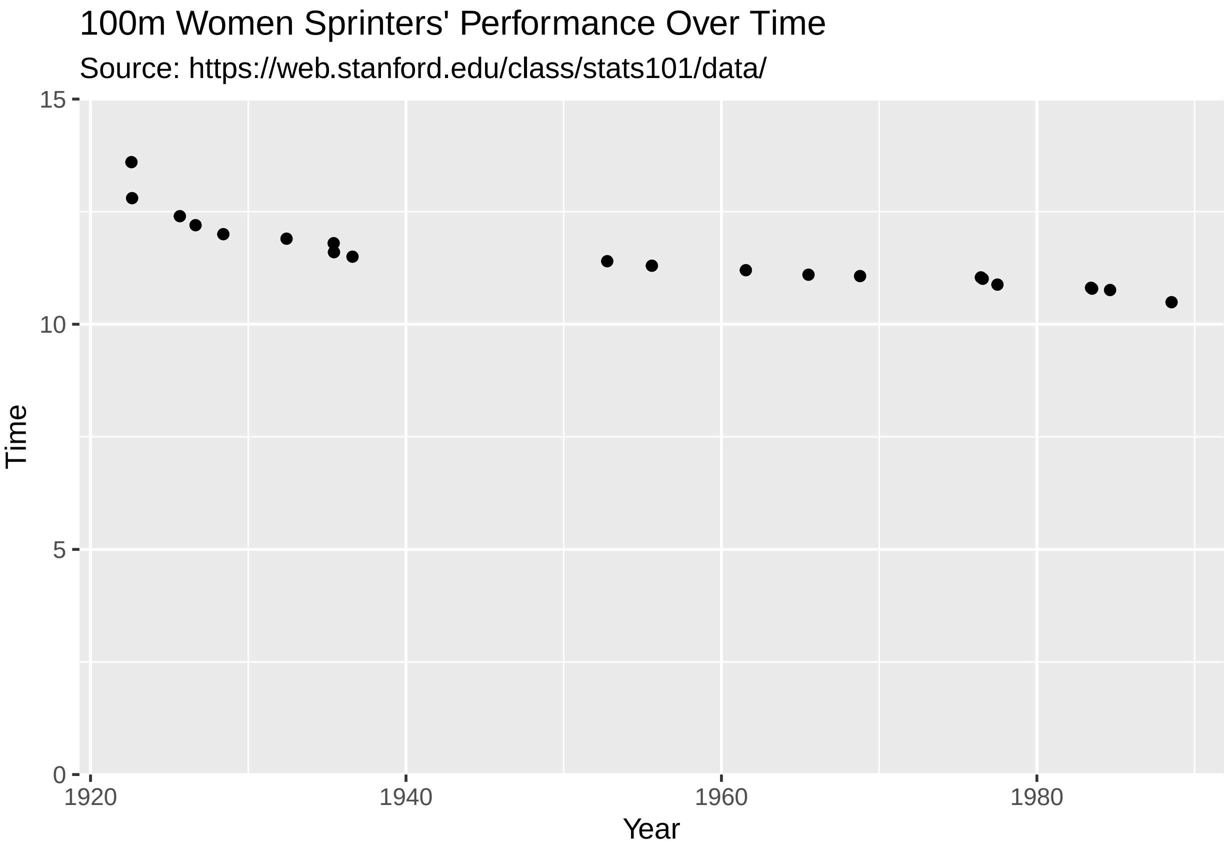
```
b_nation2 = rbind(
    data.frame(b_nation$nationality, "frequency" = b_nation$Freq.x, "type"="women"),
     data.frame(b_nation$nationality, "frequency" = b_nation$Freq.y, "type"="men"))
colnames(b_nation2)[1] = "nationalities"
ggplot(b_nation2, aes(x=nationalities, y=frequency, fill=type)) +
    geom_bar(stat="identity", position = "dodge") + scale_x_discrete(guide = guide_axis(n.dodge=3)) + g
            position = position_dodge(1), vjust = -0.1) + labs(y = "Frequency", title = "Nationalities
```


Nationalities of Men and Women Sprinters

Source: https://web.stanford.edu/class/stats101/data/

c)

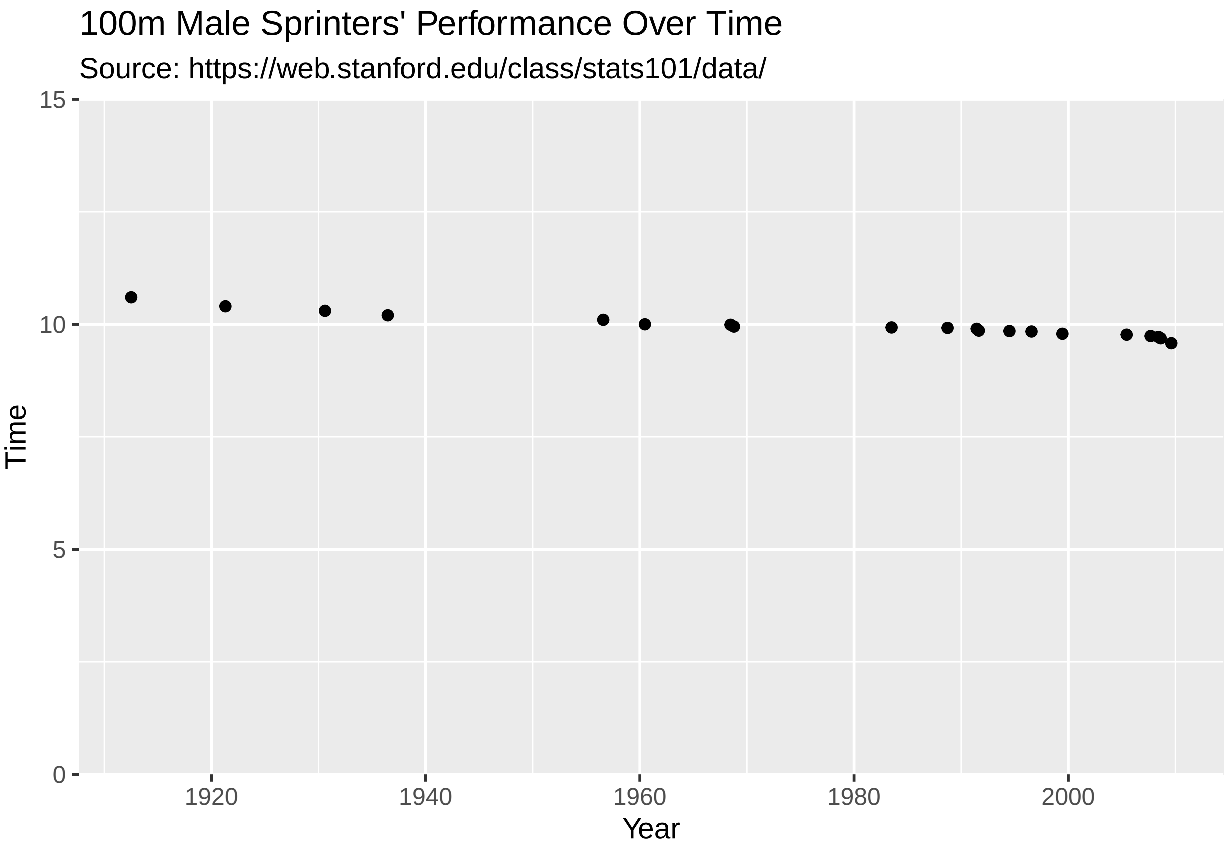
```
wPerformance_plot = ggplot(women100,aes(x=nDate,y=Time)) + scale_y_continuous(expand = c(0, 0), limits
print(wPerformance_plot)
```



100m Women Sprinters' Performance Over Time

Source: https://web.stanford.edu/class/stats101/data/

```
men100$nDate <- as.Date(men100$Date)
mPerformance_plot = ggplot(men100,aes(x=nDate,y=Time)) + scale_y_continuous(expand = c(0, 0), limits = c
print(mPerformance_plot)
```


100m Male Sprinters' Performance Over Time

Source: https://web.stanford.edu/class/stats101/data/

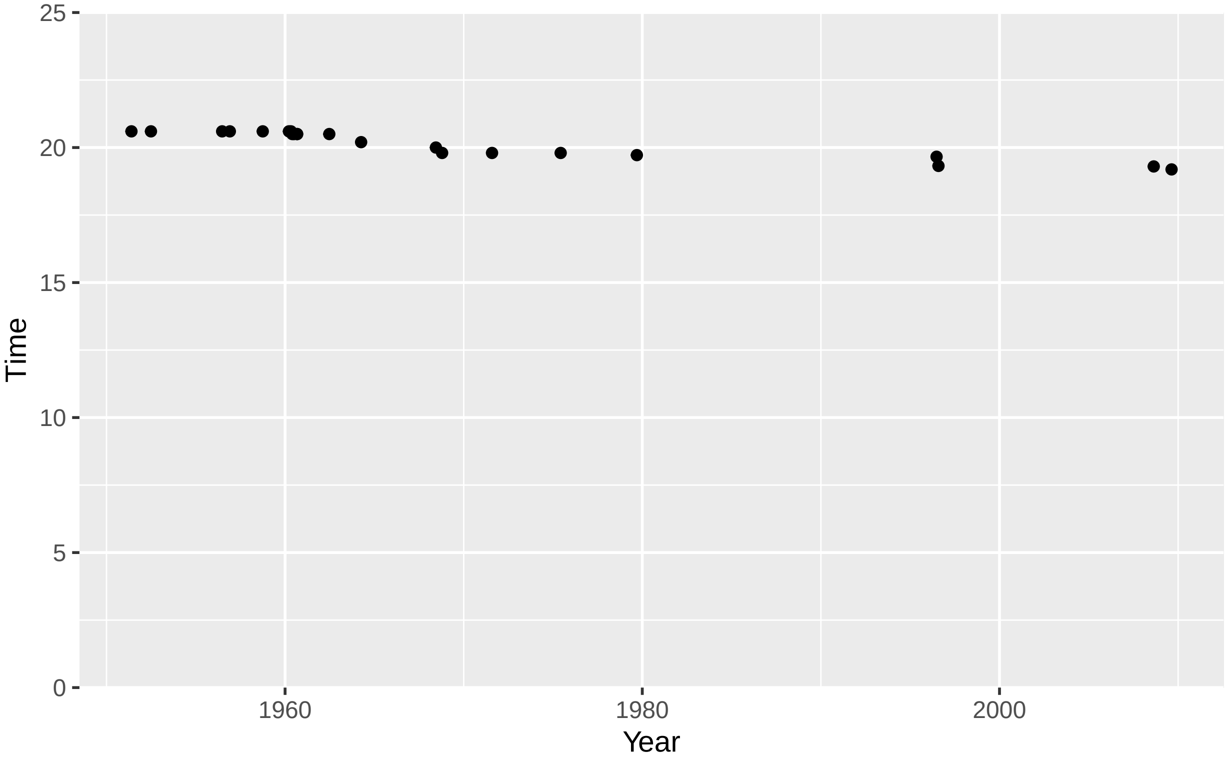
Women are improving quicker than men are. As seen in the graph showing women's performance over time, the overall change in the times of the women's run has a much larger change than with the men's. The men's graph shows the times mainly staying the same over time. It does seem women are "catching up" to the men's times, as the men's times stay around 10 seconds, and the women's times are approaching 10 seconds as time increases.

d)

```
men200 <- read.csv("./data/200men.csv",sep='\t')
men200$nDate <-  as.Date(as.character(men200$Date),format='%b %d, %Y')

ggplot(men200,aes(x=nDate,y=Time)) + scale_y_continuous(expand = c(0, 0), limits = c(0, 25)) + geom_poi
```

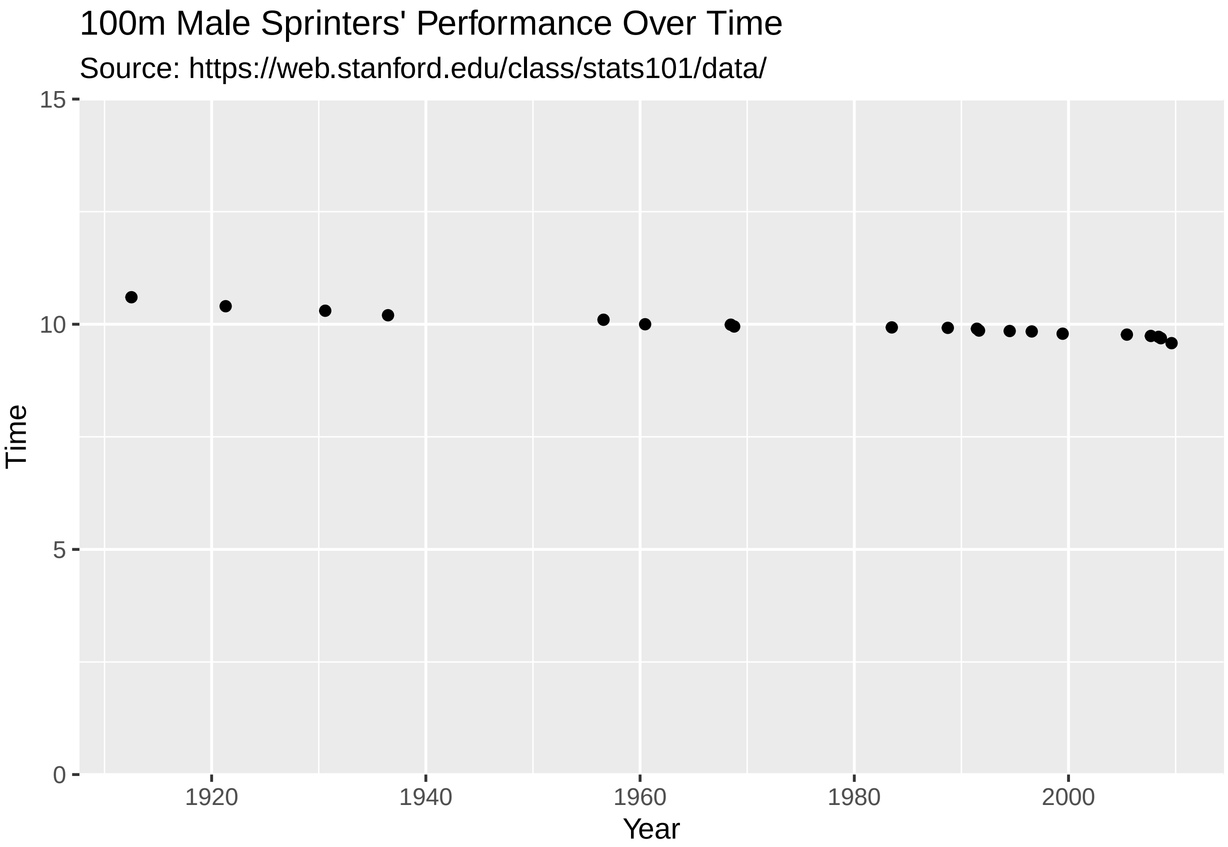
```
## Warning: Removed 3 rows containing missing values (geom_point).
```



200m Male Sprinters' Performance Over Time

Source: https://web.stanford.edu/class/stats101/data/

```
mPerformance_plot
```



100m Male Sprinters' Performance Over Time

Source: https://web.stanford.edu/class/stats101/data/

The 200m sprints are nearly twice as slow as the 100m sprints. The plot showing 200m sprints shows the average being around 20 seconds, yet the plot showing the 100m sprints shows the average being around 10s.

4) Data scientists working for Facebook and other entities that employ questionalbe ethics in theri research have been led on by a misconception referred to as the "religion of Big Data". They believe that the main focus is collecting large volumes of data, regardless of the source, and accord large sample sizes to the merit of a study. Snow did not push for more data as he carried out his invesgitation, and instead it was focused on the areas surrounding the Broad Street Pump, rather than other pumps in the U.K.