# Exploring data Assignment

Alec Brooks

2024-10-12

## Question: 1

**Q1.1: Based on this dataset, when was the first year of the Oscar Awards?**

```
summary(oscars) #get the variable name scar_yr
```

```
##     oscar_no        oscar_yr       award               name
##  Min.   : 1.00   Min.   :1929   Length:184         Length:184
##  1st Qu.:23.00   1st Qu.:1951   Class :character   Class :character
##  Median :45.50   Median :1974   Mode  :character   Mode  :character
##  Mean   :45.75   Mean   :1974
##  3rd Qu.:68.25   3rd Qu.:1996
##  Max.   :91.00   Max.   :2019
##     movie                age          birth_pl           birth_date
##  Length:184         Min.   :21.00   Length:184         Min.   :1868-04-10
##  Class :character   1st Qu.:32.00   Class :character   1st Qu.:1910-02-17
##  Mode  :character   Median :38.00   Mode  :character   Median :1932-02-27
##                     Mean   :40.05                      Mean   :1933-03-12
##                     3rd Qu.:46.00                      3rd Qu.:1956-09-20
##                     Max.   :80.00                      Max.   :1990-08-15
##     birth_mo        birth_d         birth_y
##  Min.   : 1.00   Min.   : 1.00   Min.   :1868
##  1st Qu.: 4.00   1st Qu.: 7.00   1st Qu.:1910
##  Median : 6.00   Median :15.00   Median :1932
##  Mean   : 6.44   Mean   :15.65   Mean   :1933
##  3rd Qu.: 9.00   3rd Qu.:24.00   3rd Qu.:1956
##  Max.   :12.00   Max.   :31.00   Max.   :1990
```

```
min(oscars$oscar_yr) #get the lowest value for first year
```

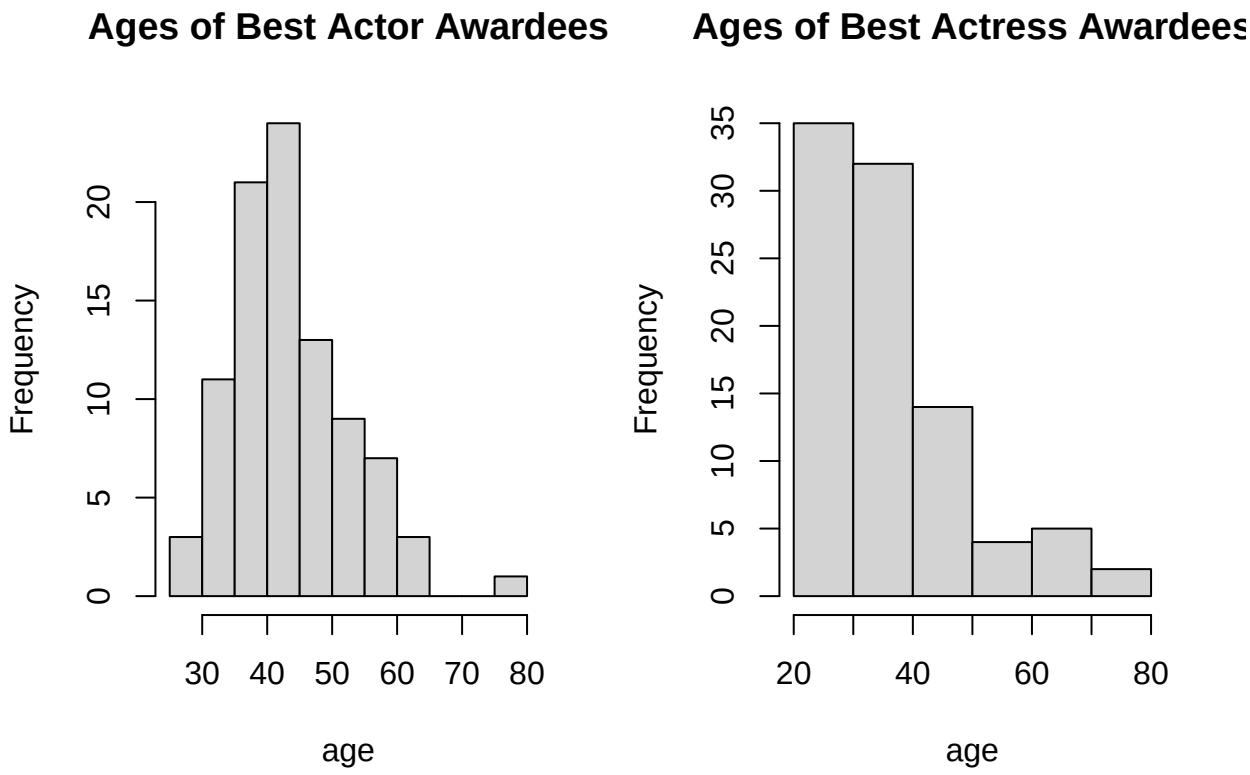
```
## [1] 1929
```

**Q1.2: Create two histograms to show the distribution of ages. One is for the best actor and the other for the best actress.**

```
par(mfrow = c(1, 2))
table(oscars$award) #see what awards are present.
```

```
##
##   Best actor Best actress
##           92           92
```

```r
hist(subset(oscars, oscars$award == "Best actor", select = age)$age, #plot subset of male
     main = "Ages of Best Actor Awardees",
     xlab = "age",
     ylab = "Frequency")
hist(subset(oscars, oscars$award == "Best actress", select = age)$age, #plot subset of female
     main = "Ages of Best Actress Awardees",
     xlab = "age",
     ylab = "Frequency")
```




**Q1.3: Find means and standard deviations of the best actors' and actresses' ages.**

```r
StatTable <- data.frame(
  Gender = c("Male", "Male", "Female", "Female"),
  Statistic = c("Mean", "SD", "Mean", "SD"),
  Value = c(mean(subset(oscars, oscars$award == "Best actor", select = age)$age),
            sd(subset(oscars, oscars$award == "Best actor", select = age)$age),
            mean(subset(oscars, oscars$award == "Best actress", select = age)$age),
            sd(subset(oscars, oscars$award == "Best actress", select = age)$age))
)
print(StatTable)
```

```
##   Gender Statistic     Value
## 1   Male      Mean 43.847826
## 2   Male        SD  8.825456
## 3 Female      Mean 36.250000
## 4 Female        SD 11.904035
```

```
rm(StatTable)
```

**Q1.4: Compare the distributions of ages of best actor and actress winners based on 1-3. Explain your analysis.**

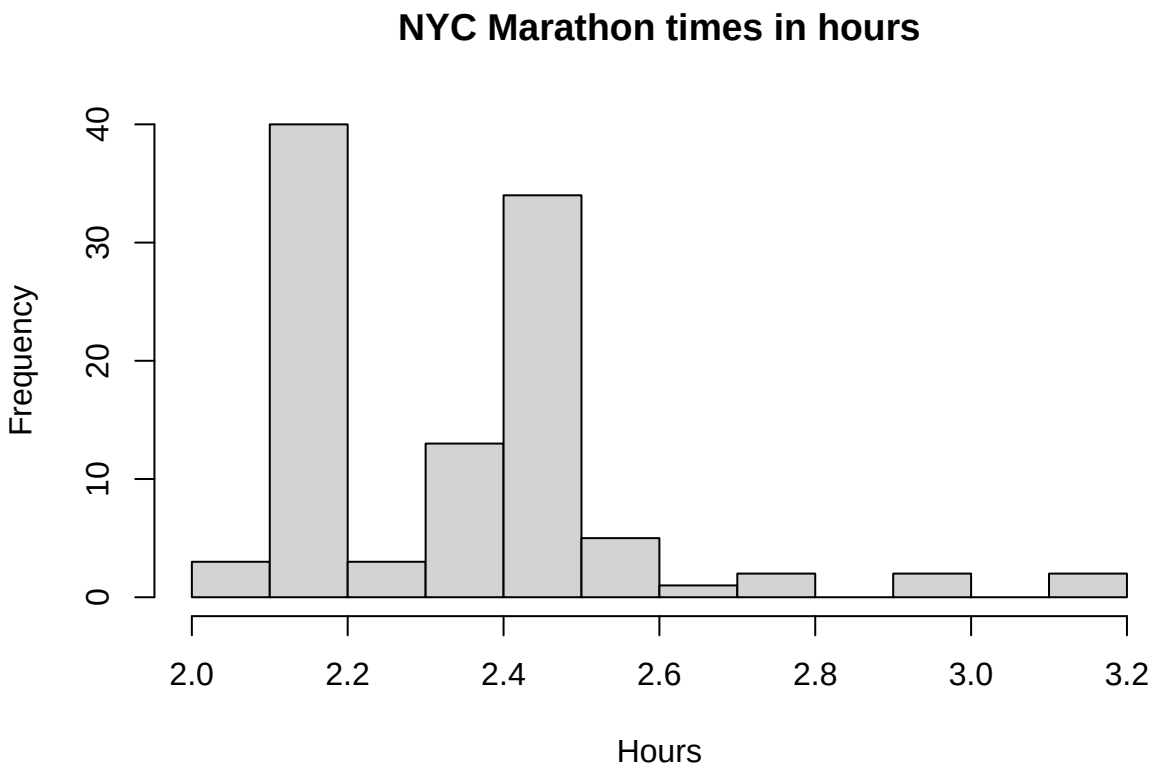
The mean age of male "Best Actor" awardees is 43.8, while for female "Best Actress" awardees, it is 36.2. This indicates that, on average, men receive the award at an older age than women. However, the standard deviation for males is lower (8.8), suggesting that the ages of male winners are more consistently closer to the mean. In contrast, the higher standard deviation for females (11.9) indicates greater variation in the ages of female awardees.

When examining the histograms, we see that both graphs are right-skewed, indicating that both cohorts are more likely to receive the award below the mean than above it. The skew for females appears more significant, suggesting that female winners falling outside the mean are more likely to be younger rather than older.

# Question: 2

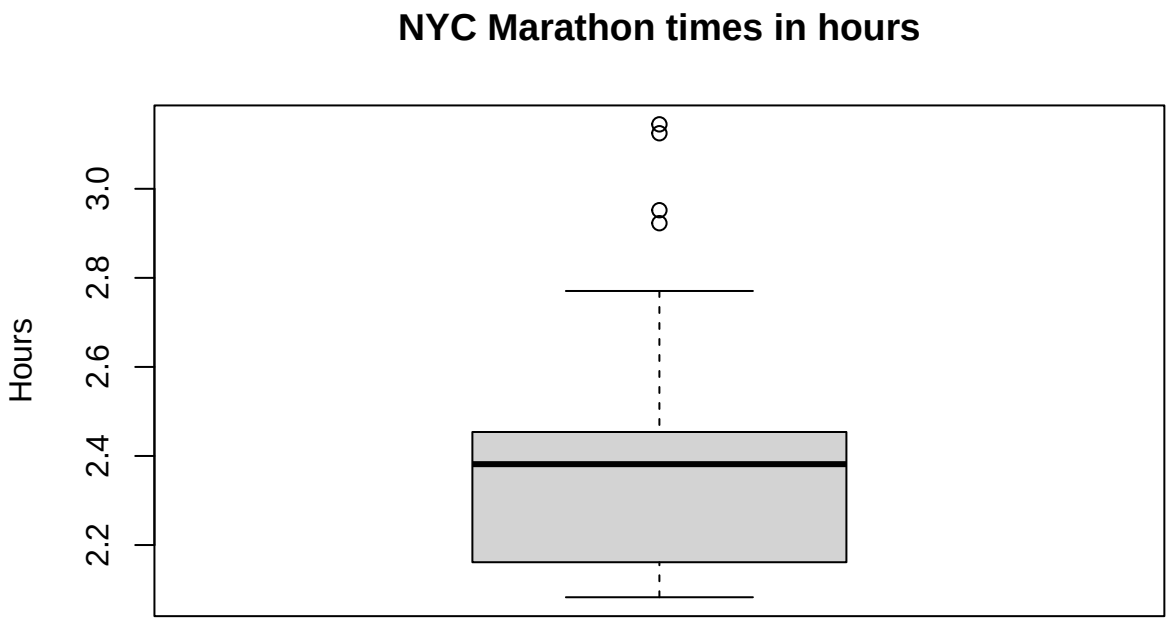
**Q2.1: Plot a histogram of time_hrs**

```
hist(nyc_marathon$time_hrs,
     main = "NYC Marathon times in hours",
     xlab = "Hours",
     ylab = "Frequency")
```




**Q2.2: Plot a boxplot of time_hrs**

```
boxplot(nyc_marathon$time_hrs,
     main = "NYC Marathon times in hours",
     xlab = "",
     ylab = "Hours")
```

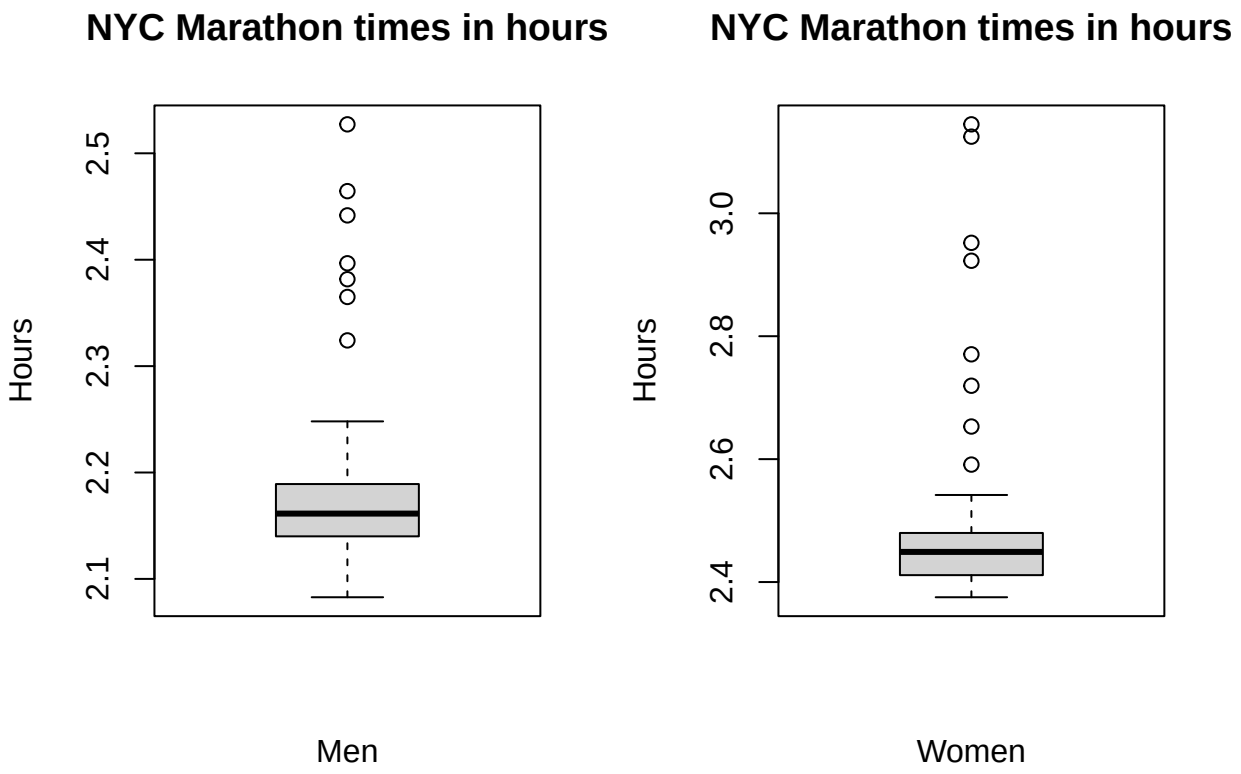



**Q2.3: Based on the graphs from 1 and 2, what features of the distribution are apparent in the histogram and not the box plot? What features are apparent in the box plot but not in the histogram?**

Both charts show the right skew of the data, with the histogram displaying it through its tail and the box plot through its whiskers. The histogram also reveals that the data is bimodal, with significant peaks between 2.1–2.2 and 2.4–2.5. It also shows these modal peaks are significant compared to the rest of the data. On the other hand the box plot, identifies four outliers well outside the range of the first (Q1) and third quartiles (Q3).

**Q2.4: Plot a boxplot of marathon times for men and women.**

```
par(mfrow = c(1, 2))
boxplot(subset(nyc_marathon, nyc_marathon$division == "Men", select = time_hrs)$time_hrs,
     main = "NYC Marathon times in hours",
     xlab = "Men",
     ylab = "Hours")
boxplot(subset(nyc_marathon, nyc_marathon$division == "Women", select = time_hrs)$time_hrs,
        main = "NYC Marathon times in hours",
        xlab = "Women",
        ylab = "Hours")
```




**Q2.5: Compare the distribution of marathon times for men and women based on the box plot from 4. Explain.**

Comparing the two, we can see that the male plot has a significantly lower median compared to the female plot, as the center boxplot line is positioned lower along the y-axis than that of the women's plot. Additionally, the whiskers appear shorter on the female plot, indicating that the female data is less spread out than the male data. Both graphs show several outliers above the third quartile (Q3).

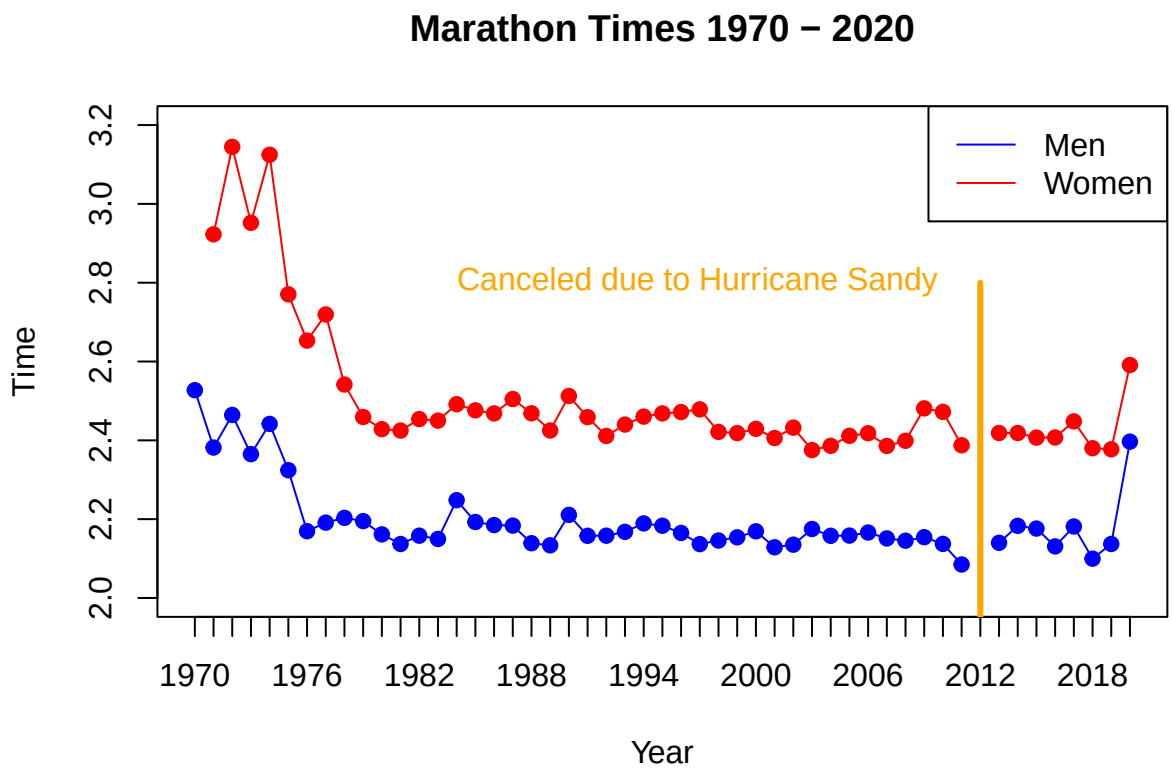
**Q2.6: Considering all the 3 graphs so far, what may be the reason for the bimodal distribution? Explain.**

The bimodal nature of the histogram is due to the two distinct means of the groups: men with a mean of approximately 2.15 hours and women at 2.5 hours. This clear distinction, combined with the fact that the histogram represents both groups together, explains why it is bimodal.

**Q2.7: Create a time series plot comparing men's and women's marathon times between 1970 and 2020. Try to add a legend here!**

```
plot(subset(nyc_marathon, nyc_marathon$division == "Men" & nyc_marathon$year >=
              1970 & nyc_marathon$year <= 2020)$year,
     subset(nyc_marathon, nyc_marathon$division == "Men" & nyc_marathon$year >=
              1970 & nyc_marathon$year <= 2020)$time_hrs,
     type = "p", pch = 21, col = "blue", bg = "blue",
```

```
     main ="Marathon Times 1970 - 2020", xlab = "Year",
     ylab = "Time", ylim = c(2,3.2), xaxt = "n")
axis(1, at = seq(1970, 2020, by = 1))
points(subset(nyc_marathon, nyc_marathon$division == "Women" & nyc_marathon$year >=
               1970 & nyc_marathon$year <= 2020)$year,
      subset(nyc_marathon, nyc_marathon$division == "Women" & nyc_marathon$year >=
               1970 & nyc_marathon$year <= 2020)$time_hrs,
      type = "p", pch = 21, col = "red", bg = "red",)
lines(subset(nyc_marathon, nyc_marathon$division == "Women" & nyc_marathon$year >=
               1970 & nyc_marathon$year <= 2020)$year,
     subset(nyc_marathon, nyc_marathon$division == "Women" & nyc_marathon$year >=
              1970 & nyc_marathon$year <= 2020)$time_hrs,
     col = "red")
lines(subset(nyc_marathon, nyc_marathon$division == "Men" & nyc_marathon$year >=
               1970 & nyc_marathon$year <= 2020)$year,
      subset(nyc_marathon, nyc_marathon$division == "Men" & nyc_marathon$year >=
               1970 & nyc_marathon$year <= 2020)$time_hrs,
      col = "blue")
legend("topright", legend = c("Men", "Women"), col = c("blue", "red"), lty = 1)
lines(c(2012, 2012), c(1.9, 2.8), col = "orange", lwd = 3)
text(1983, 2.8, labels =nyc_marathon[nyc_marathon$year ==
      2012 & nyc_marathon$division == "Men", 7], pos = 4, col = "orange")
```



**Q2.8: Describe what is visible in this plot from 7 but not in the others.**

This chart offers some unique insights. Firstly, we can clearly observe the gap between male and female performance, as well as the significant progress both cohorts made, especially between the years 1970 and 1975. Additionally, we can see the absence of data in the year 2012, which is noted in the dataset due to the marathon's cancellation. Lastly, towards the tail end of the time series, we see a noticeable increase in times for both cohorts. This may be attributed to the effects of COVID-19, a period when marathon runners might not have been operating at their peak. However, more evidence would be needed to confirm this as the cause.

# full Code

```
#install.packages("openintro")
library(openintro)
options(scipen = 999) # make it stop using scientific - google
summary(oscars) #get the variable name scar_yr
min(oscars$oscar_yr) #get the lowest value for first year
par(mfrow = c(1, 2))
table(oscars$award) #see what awards are present.
hist(subset(oscars, oscars$award == "Best actor", select = age)$age, #plot subset of male
     main = "Ages of Best Actor Awardees",
     xlab = "age",
     ylab = "Frequency")
hist(subset(oscars, oscars$award == "Best actress", select = age)$age, #plot subset of female
     main = "Ages of Best Actress Awardees",
     xlab = "age",
     ylab = "Frequency")
StatTable <- data.frame(
  Gender = c("Male", "Male", "Female", "Female"),
  Statistic = c("Mean", "SD", "Mean", "SD"),
  Value = c(mean(subset(oscars, oscars$award == "Best actor", select = age)$age),
            sd(subset(oscars, oscars$award == "Best actor", select = age)$age),
            mean(subset(oscars, oscars$award == "Best actress", select = age)$age),
            sd(subset(oscars, oscars$award == "Best actress", select = age)$age))
)
print(StatTable)
rm(StatTable)
hist(nyc_marathon$time_hrs,
     main = "NYC Marathon times in hours",
     xlab = "Hours",
     ylab = "Frequency")
boxplot(nyc_marathon$time_hrs,
        main = "NYC Marathon times in hours",
        xlab = "",
        ylab = "Hours")
par(mfrow = c(1, 2))
boxplot(subset(nyc_marathon, nyc_marathon$division == "Men", select = time_hrs)$time_hrs,
        main = "NYC Marathon times in hours",
        xlab = "Men",
        ylab = "Hours")
boxplot(subset(nyc_marathon, nyc_marathon$division == "Women", select = time_hrs)$time_hrs,
```

```
        main = "NYC Marathon times in hours",
        xlab = "Women",
        ylab = "Hours")
plot(subset(nyc_marathon, nyc_marathon$division == "Men" & nyc_marathon$year >=
               1970 & nyc_marathon$year <= 2020)$year,
     subset(nyc_marathon, nyc_marathon$division == "Men" & nyc_marathon$year >=
               1970 & nyc_marathon$year <= 2020)$time_hrs,
     type = "p", pch = 21, col = "blue", bg = "blue",
     main ="Marathon Times 1970 - 2020", xlab = "Year",
     ylab = "Time", ylim = c(2,3.2), xaxt = "n")
axis(1, at = seq(1970, 2020, by = 1))
points(subset(nyc_marathon, nyc_marathon$division == "Women" & nyc_marathon$year >=
                1970 & nyc_marathon$year <= 2020)$year,
       subset(nyc_marathon, nyc_marathon$division == "Women" & nyc_marathon$year >=
                1970 & nyc_marathon$year <= 2020)$time_hrs,
       type = "p", pch = 21, col = "red", bg = "red",)
lines(subset(nyc_marathon, nyc_marathon$division == "Women" & nyc_marathon$year >=
               1970 & nyc_marathon$year <= 2020)$year,
      subset(nyc_marathon, nyc_marathon$division == "Women" & nyc_marathon$year >=
               1970 & nyc_marathon$year <= 2020)$time_hrs,
      col = "red")
lines(subset(nyc_marathon, nyc_marathon$division == "Men" & nyc_marathon$year >=
               1970 & nyc_marathon$year <= 2020)$year,
      subset(nyc_marathon, nyc_marathon$division == "Men" & nyc_marathon$year >=
               1970 & nyc_marathon$year <= 2020)$time_hrs,
      col = "blue")
legend("topright", legend = c("Men", "Women"), col = c("blue", "red"), lty = 1)
lines(c(2012, 2012), c(1.9, 2.8), col = "orange", lwd = 3)
text(1983, 2.8, labels =nyc_marathon[nyc_marathon$year ==
      2012 & nyc_marathon$division == "Men", 7], pos = 4, col = "orange")
```