HW MD 1 and 2

Emilio

2022-09-02

1.

```
install.packages("causact")
```

```
## Installing package into '/home/guest/R/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages("dplyr")
```

```
## Installing package into '/home/guest/R/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages("igraph")
```

```
## Installing package into '/home/guest/R/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(causact)
library (dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(igraph)
```

```
##
## Attaching package: 'igraph'
```

```
## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union
```

```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
```

```
## The following object is masked from 'package:base':
##
##     union
```

2.

the error message it gives says "Error in as_data_frame(x = c(1, 2, 3)) : Not a graph object"

```
df <- dplyr::as_data_frame(x = c(1,2,3))
```

```
## Warning: `as_data_frame()` was deprecated in tibble 2.0.0.
## Please use `as_tibble()` instead.
## The signature and semantics have changed, see `?as_tibble`.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

```
glimpse(df)
```

```
## Rows: 3
## Columns: 1
## $ value <dbl> 1, 2, 3
```

```
x <- c(5,6,2,7,9,1)
dplyr::n_distinct(x)
```

```
## [1] 6
```

The order in which you load packages matters. If they have identital functions with identical names R draws from the one loaded last

3.

```
?n_distinct
```

N distinct essentially is a contraction of length(unique(x)). It counts the number of unique values in a set of vectors.

4.

```
glimpse(baseballData)
```

```
## Rows: 12,145
## Columns: 5
## $ Date         <int> 20100405, 20100405, 20100405, 20100405, 20100405, 2010040~
## $ Home         <fct> ANA, CHA, KCA, OAK, TEX, ARI, ATL, CIN, HOU, MIL, NYN, PI~
## $ Visitor      <fct> MIN, CLE, DET, SEA, TOR, SDN, CHN, SLN, SFN, COL, FLO, LA~
## $ HomeScore    <int> 6, 6, 4, 3, 5, 6, 16, 6, 2, 3, 7, 11, 1, 3, 4, 2, 4, 3, 0~
## $ VisitorScore <int> 3, 0, 8, 5, 4, 3, 5, 11, 5, 5, 1, 5, 11, 5, 6, 1, 3, 6, 3~
```

rows: 12,145 columns: 5

Home is an fct variable while HomeScore is an int

```
baseballData[1,]
```

```
##       Date Home Visitor HomeScore VisitorScore
## 1 20100405  ANA     MIN         6            3
```

```
?tibble
```

```
## Help on topic 'tibble' was found in the following packages:
##
##   Package               Library
##   tidyr                 /usr/local/lib/R/site-library
##   tibble                /usr/local/lib/R/site-library
##   dplyr                 /home/guest/R/x86_64-pc-linux-gnu-library/4.2
##
##
```

```
## Using the first match ...
```

```
baseballData[,2:3] %>% head()
```

```
##   Home Visitor
## 1  ANA     MIN
## 2  CHA     CLE
## 3  KCA     DET
## 4  OAK     SEA
## 5  TEX     TOR
## 6  ARI     SDN
```

The row represents one game in the dataset. The column represents which team was home and which team was away.

6.

```
name <-
  c("Wayne Gretzky",
    "Gordie Howe",
    "Jaromir Jagr",
    "Brett Hull",
    "Marcel Dionne",
    "Phil Esposito" ,
    "Mike Gartner",
    "Alex Ovechkin",
    "Mark Messier" ,
    "Steve Yzerman"
  )

goals <- c(894, 801, 766, 741, 731, 717, 708, 700, 694, 692)

year_started <- c(1979, 1946, 1990, 1986, 1971, 1963, 1979, 2005, 1979, 1983)
```

```
df <- tibble(
  name = name,
  goals = goals,
  year_started = year_started)
```

```
glimpse (df)
```

```
## Rows: 10
## Columns: 3
## $ name         <chr> "Wayne Gretzky", "Gordie Howe", "Jaromir Jagr", "Brett Hu~
## $ goals        <dbl> 894, 801, 766, 741, 731, 717, 708, 700, 694, 692
## $ year_started <dbl> 1979, 1946, 1990, 1986, 1971, 1963, 1979, 2005, 1979, 1983
```

MD chapter 2

```
library(tidyverse)
```

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v stringr 1.4.1
## v tidyr   1.2.0     v forcats 0.5.2
## v readr   2.1.2
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x tibble::as_data_frame() masks igraph::as_data_frame(), dplyr::as_data_frame()
## x purrr::compose()        masks igraph::compose()
## x tidyr::crossing()       masks igraph::crossing()
## x dplyr::filter()         masks stats::filter()
## x igraph::groups()        masks dplyr::groups()
## x dplyr::lag()            masks stats::lag()
## x purrr::simplify()       masks igraph::simplify()
```

```
olympics  <- read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2
```

```
## Rows: 271116 Columns: 15
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (10): name, sex, team, noc, games, season, city, sport, event, medal
## dbl  (5): id, age, height, weight, year
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(olympics)
```

```
## Rows: 271,116
## Columns: 15
## $ id     <dbl> 1, 2, 3, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 6, 7, 7, 7, ~
## $ name   <chr> "A Dijiang", "A Lamusi", "Gunnar Nielsen Aaby", "Edgar Lindenau~
## $ sex    <chr> "M", "M", "M", "M", "F", "F", "F", "F", "F", "F", "M", "M", "M"~
## $ age    <dbl> 24, 23, 24, 34, 21, 21, 25, 25, 27, 27, 31, 31, 31, 31, 33, 33,~
## $ height <dbl> 180, 170, NA, NA, 185, 185, 185, 185, 185, 185, 188, 188, 188, ~
## $ weight <dbl> 80, 60, NA, NA, 82, 82, 82, 82, 82, 82, 75, 75, 75, 75, 75, 75,~
## $ team   <chr> "China", "China", "Denmark", "Denmark/Sweden", "Netherlands", "~
## $ noc    <chr> "CHN", "CHN", "DEN", "DEN", "NED", "NED", "NED", "NED", "NED", ~
## $ games  <chr> "1992 Summer", "2012 Summer", "1920 Summer", "1900 Summer", "19~
## $ year   <dbl> 1992, 2012, 1920, 1900, 1988, 1988, 1992, 1992, 1994, 1994, 199~
## $ season <chr> "Summer", "Summer", "Summer", "Summer", "Winter", "Winter", "Wi~
## $ city   <chr> "Barcelona", "London", "Antwerpen", "Paris", "Calgary", "Calgar~
## $ sport  <chr> "Basketball", "Judo", "Football", "Tug-Of-War", "Speed Skating"~
## $ event  <chr> "Basketball Men's Basketball", "Judo Men's Extra-Lightweight", ~
## $ medal  <chr> NA, NA, NA, "Gold", NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
```

```
table(olympics$medal)
```

```
##
## Bronze   Gold Silver
##  13295  13372  13116
```

```
gold_medalists <- olympics %>%
  filter(medal == "Gold")
glimpse(gold_medalists)
```

```
## Rows: 13,372
## Columns: 15
## $ id     <dbl> 4, 17, 17, 17, 20, 20, 20, 20, 21, 40, 42, 56, 72, 73, 73, 76, ~
## $ name   <chr> "Edgar Lindenau Aabye", "Paavo Johannes Aaltonen", "Paavo Johan~
## $ sex    <chr> "M", "M", "M", "M", "M", "M", "M", "M", "F", "M", "M", "M", "M"~
## $ age    <dbl> 34, 28, 28, 28, 20, 30, 30, 34, 27, 31, 25, 21, 28, 23, 27, 22,~
## $ height <dbl> NA, 175, 175, 175, 176, 176, 176, 176, 163, NA, NA, NA, 180, 18~
```

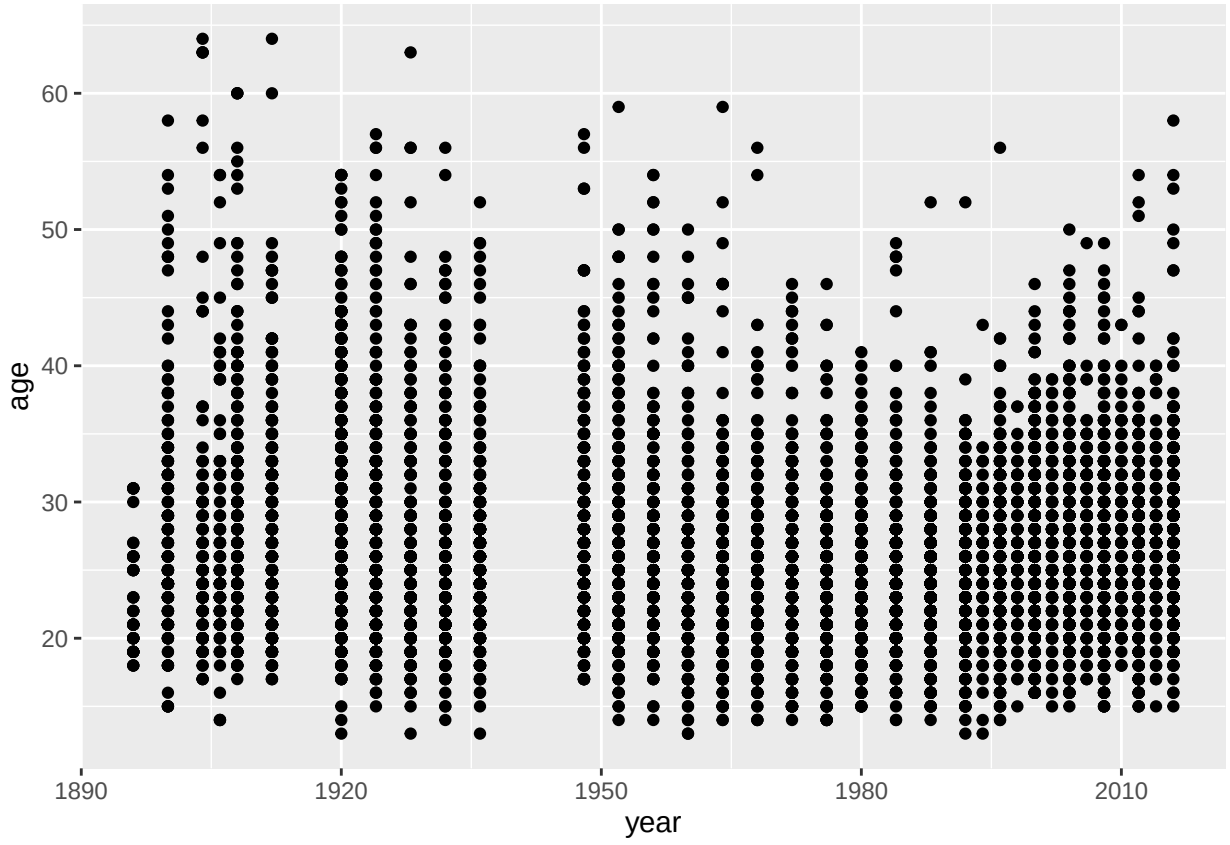
```
## $ weight <dbl> NA, 64, 64, 64, 85, 85, 85, 85, NA, NA, NA, NA, 83, 86, 86, 82,~
## $ team   <chr> "Denmark/Sweden", "Finland", "Finland", "Finland", "Norway", "N~
## $ noc    <chr> "DEN", "FIN", "FIN", "FIN", "NOR", "NOR", "NOR", "NOR", "NOR", ~
## $ games  <chr> "1900 Summer", "1948 Summer", "1948 Summer", "1948 Summer", "19~
## $ year   <dbl> 1900, 1948, 1948, 1948, 1992, 2002, 2002, 2006, 2008, 1960, 191~
## $ season <chr> "Summer", "Summer", "Summer", "Summer", "Winter", "Winter", "Wi~
## $ city   <chr> "Paris", "London", "London", "London", "Albertville", "Salt Lak~
## $ sport  <chr> "Tug-Of-War", "Gymnastics", "Gymnastics", "Gymnastics", "Alpine~
## $ event  <chr> "Tug-Of-War Men's Tug-Of-War", "Gymnastics Men's Team All-Aroun~
## $ medal  <chr> "Gold", "Gold", "Gold", "Gold", "Gold", "Gold", "Gold", "Gold",~
```

Rows: 13372

2. linegraph, x axis year y axis age

```
ggplot(data = gold_medalists,
       mapping = aes(x = year, y = age)) +
  geom_point()
```

```
## Warning: Removed 148 rows containing missing values (geom_point).
```



for overplotting you can adjust the transparency of the points or add little "jitters" to the points

3.

```
us_medals <- gold_medalists %>%
  filter(noc == "USA") %>%
  group_by(year) %>%
  summarise(num_medals = n())
```

```
ggplot(data = us_medals,
       mapping = aes(x = year, y = num_medals)) +
  geom_line()
```

The US had the most medals in 1984, but the Soviet Union protested the Olympics that year to that contributed to the high medal count.

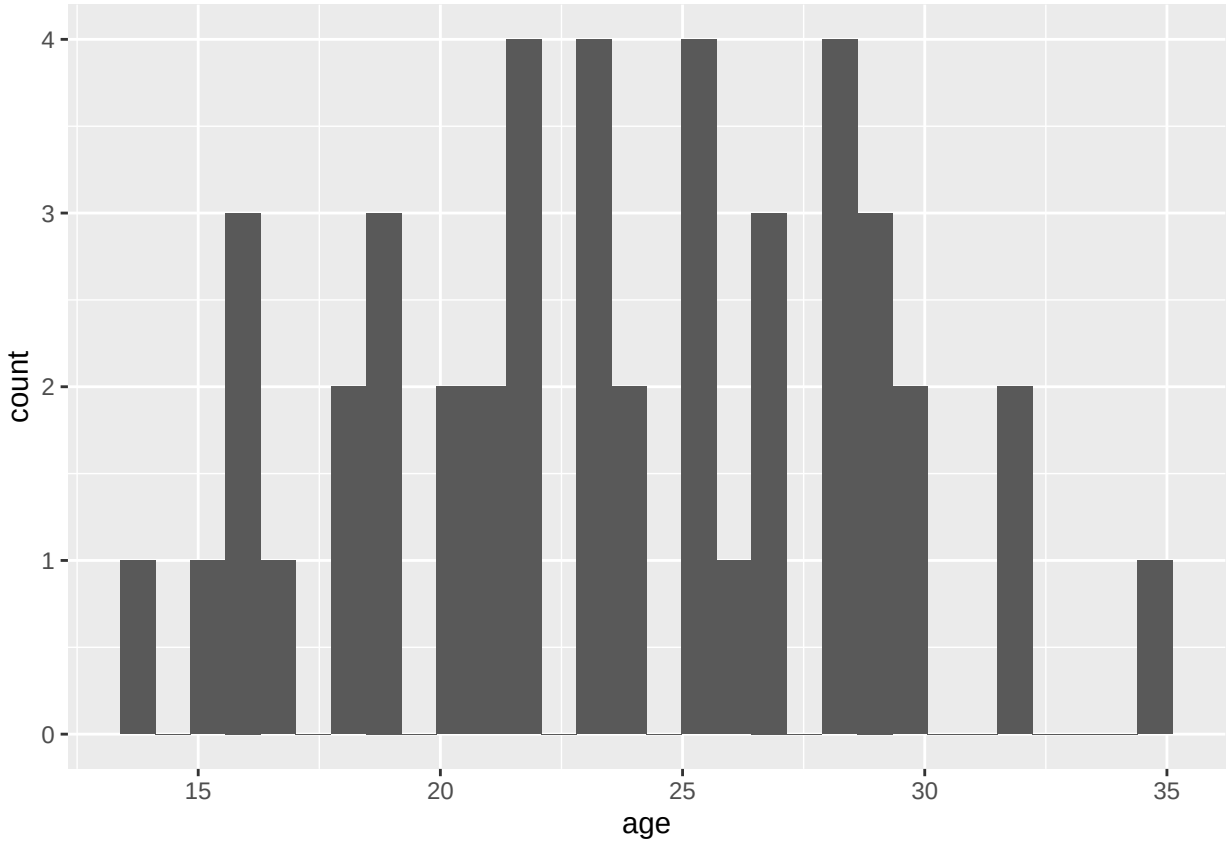
4.

```
two_events <- gold_medalists %>%
  filter(
    event == "Gymnastics Men's Individual All-Around"  |
      event == "Gymnastics Women's Individual All-Around" |
      event == "Athletics Women's 100 metres"   |
      event == "Athletics Men's 100 metres"
  )
```

```
two_events <- gold_medalists %>%
  filter(
    event == "Gymnastics Men's Individual All-Around"  |
      event == "Gymnastics Women's Individual All-Around"
)
```

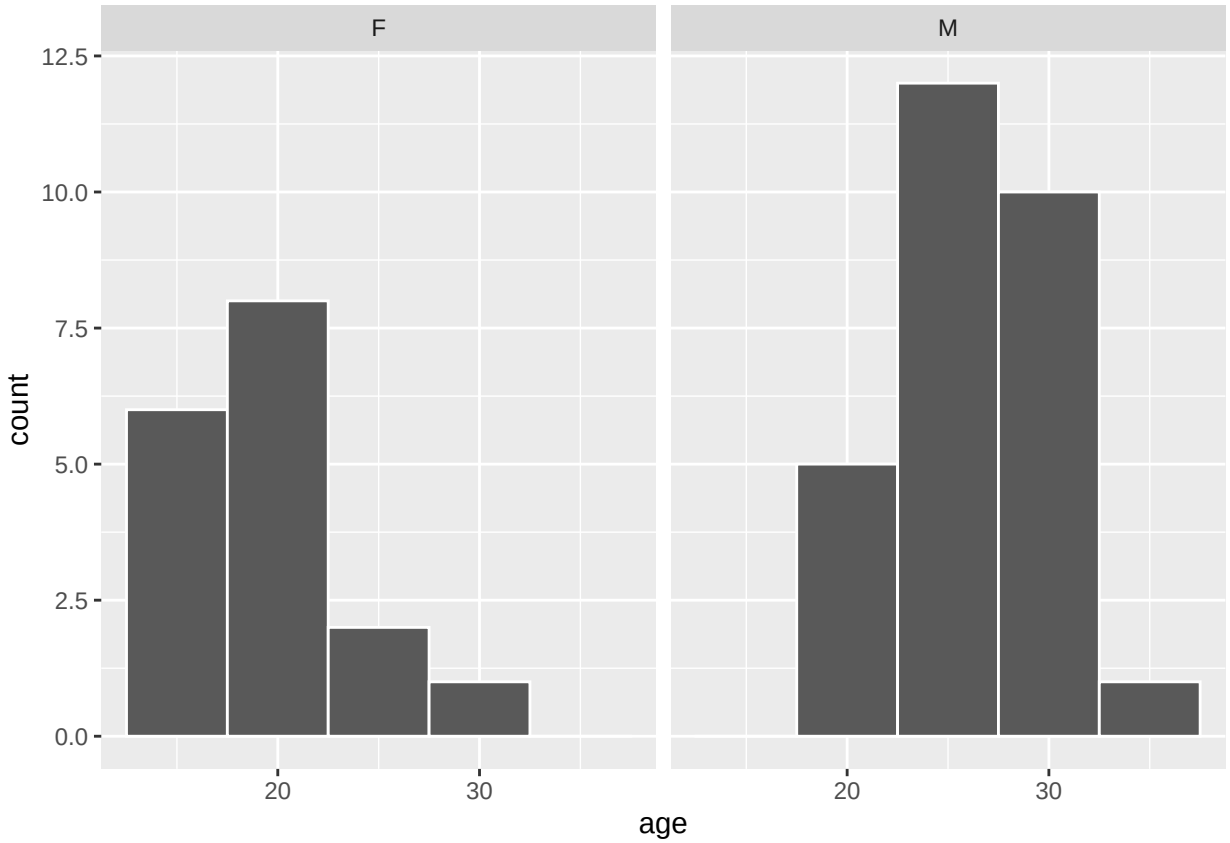
```
ggplot(data = two_events,
       mapping = aes(x = age,)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The graph describes the age distribution for US Olympic Men and Women's gymnastics team gold medalists. The range of ages is 13-34. The modal ages are 22, 23, 25, and 26.

```
ggplot(data = two_events, mapping = aes(x = age)) +
  geom_histogram(binwidth = 5, color = "white") +
  facet_wrap(~sex)
```

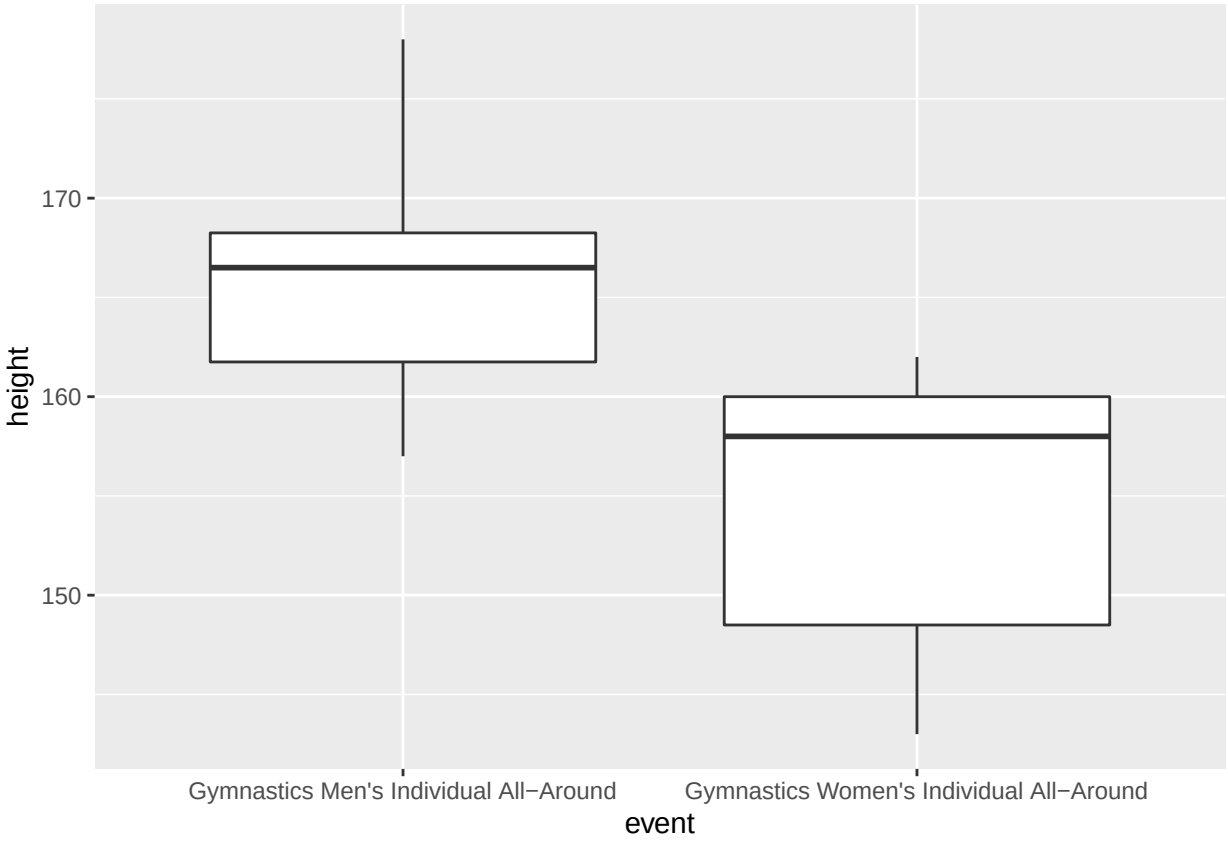


Men tend to be older

5.

```
ggplot(data = two_events, mapping = aes(x = event , y = height )) +
  geom_boxplot()
```

```
## Warning: Removed 10 rows containing non-finite values (stat_boxplot).
```

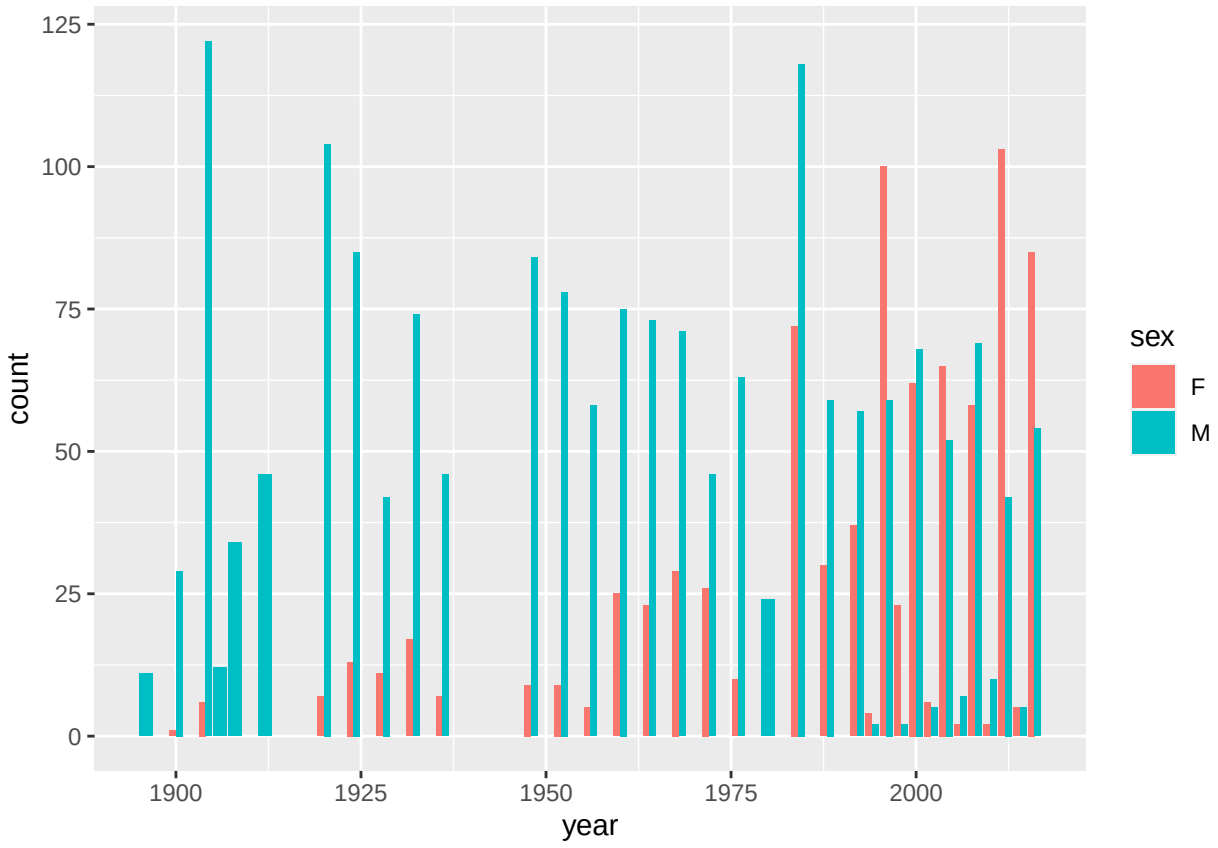


This boxplot shows that men tend to be around 166 cm while women tend to be around 158 cm.

6.

```
us_medalists <- gold_medalists %>%
  filter(noc == "USA")
```

```
ggplot(data = us_medalists, mapping = aes(x = year, fill= sex)) +
  geom_bar(position= "dodge")
```



The graph shows that recently women have been winning more medals than men for the United States.