MD_Chap1-2

Rachel Kaufman

2022-09-01

Chapter1 #Question 1

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

#Question 2

```
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
```

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

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

```
## [1] 3
```

I believe that because we attached igraph most recently (in line 20) it is masking dplyr because that one was downloaded prior to. the computer is just using the packages in order of download

#Question 3

```
?n_distinct()
```

This argument counts the number of unique values in a set of vectors, this would mean that if a column has any repetitive observations, those values would count as one. I could see this being helpful with geographic data and being able to group say by states as elements.

#Question 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~
```

```
class(baseballData$Home)
```

```
## [1] "factor"
```

```
class(baseballData$HomeScore)
```

```
## [1] "integer"
```

This dataset has 12,145 rows and 5 columns. The home column is a factor variable, and the HomeScore is an integer variable.

#Question 5

```
baseballData[1,]
```

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

```
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 data in row 1 represents a row of a single observation seen in line 49. The code in like 50 shows us the top part (i.e. the head) of the dataframe, including only 6 observations. The columns "home" and "visitor" are used because the arguement was passed for columns 2:3.

#Question 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(
  first_var_name = name,
  second_var_name = goals,
  third_var_name = year_started
  )
glimpse(df)
```

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

CHAPTER 2 Question 1:

```
library(tidyverse)
```

```
## -- 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_medalist <- olympics %>%
  filter(medal == "Gold")
glimpse(gold_medalist)
```

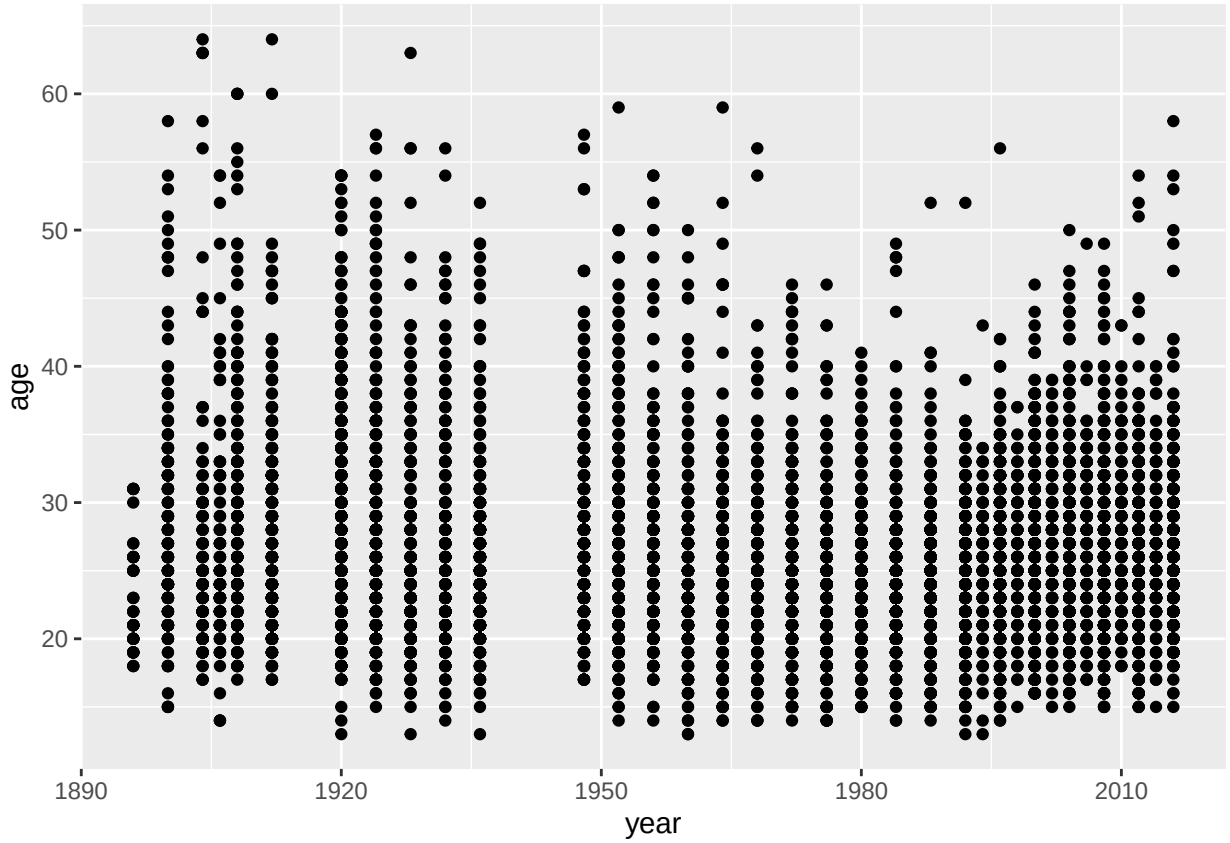
```
## Rows: 13,372
## Columns: 15
## $ id     <dbl> 4, 17, 17, 17, 20, 20, 20, 20, 21, 40, 42, 56, 72, 73, 73, 76, ~
## $ name   <chr> "Edgar Lindenau Aabye", "Paavo Johannes Aaltonen", "Paavo Johan~
## $ sex    <chr> "M", "M", "M", "M", "M", "M", "M", "M", "F", "M", "M", "M", "M"~
## $ age    <dbl> 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",~
```

There are now 13,372 rows once you filter for only Gold medalists.

Question 2:

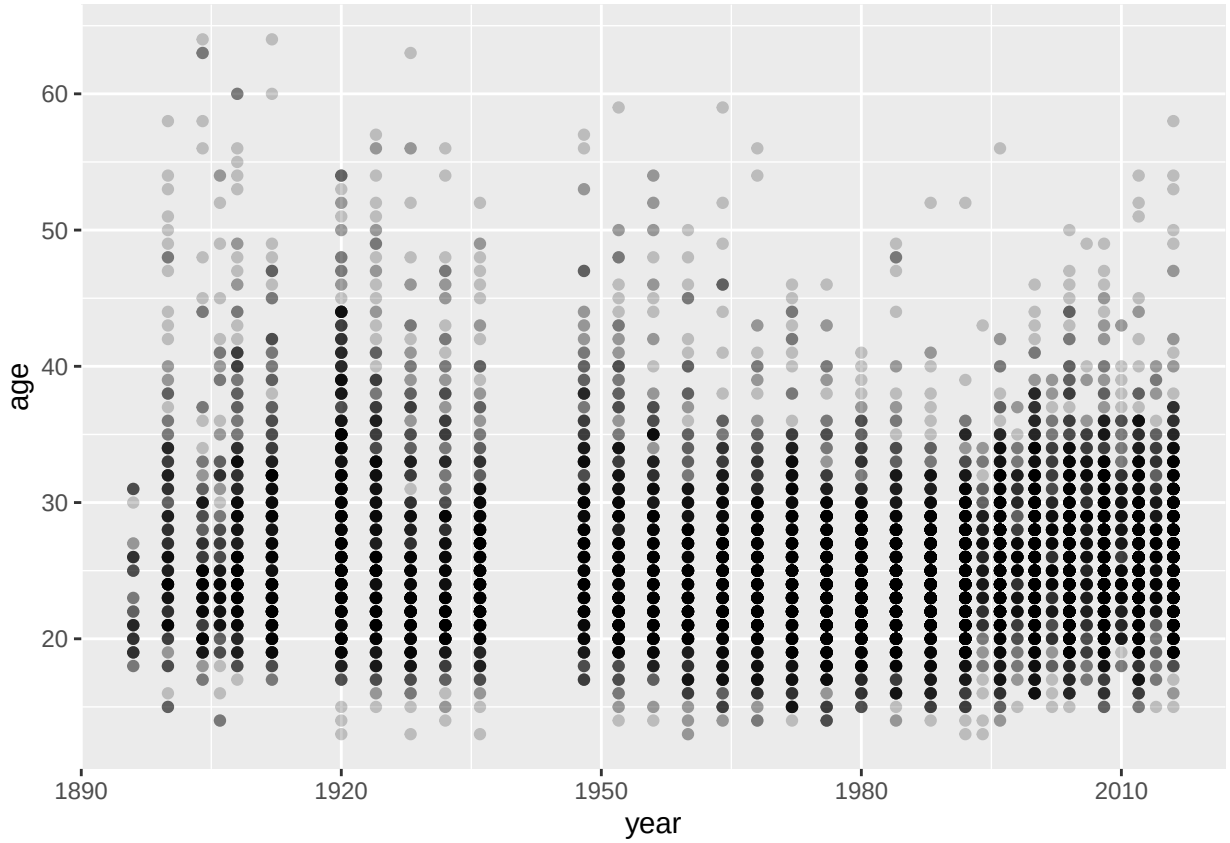
```
ggplot(gold_medalist, aes(year,age)) +
  geom_point()
```

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



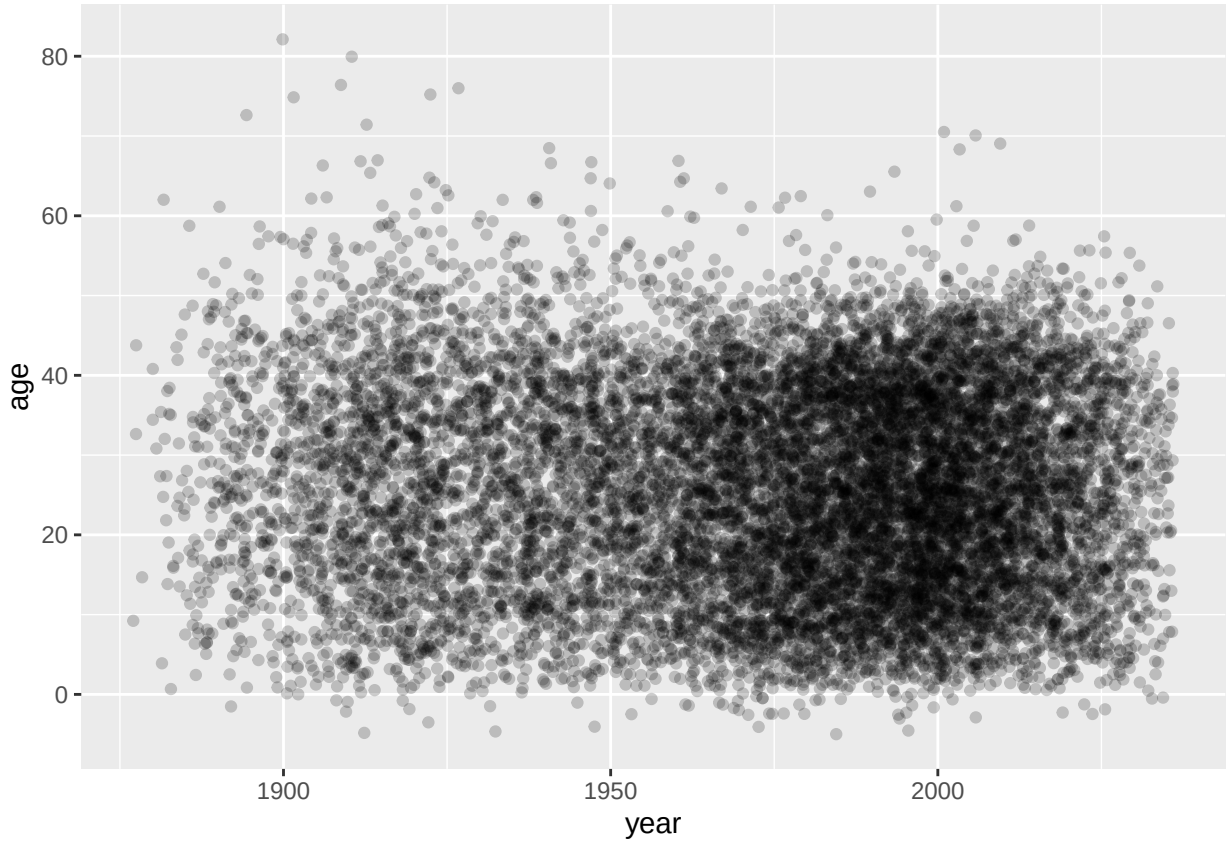

```
ggplot(gold_medalist, aes(year,age)) +
  geom_point(alpha = 0.2)
```

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




```
ggplot(gold_medalist, aes(year,age)) +
  geom_jitter(alpha = 0.2, width = 20, height = 20)
```

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



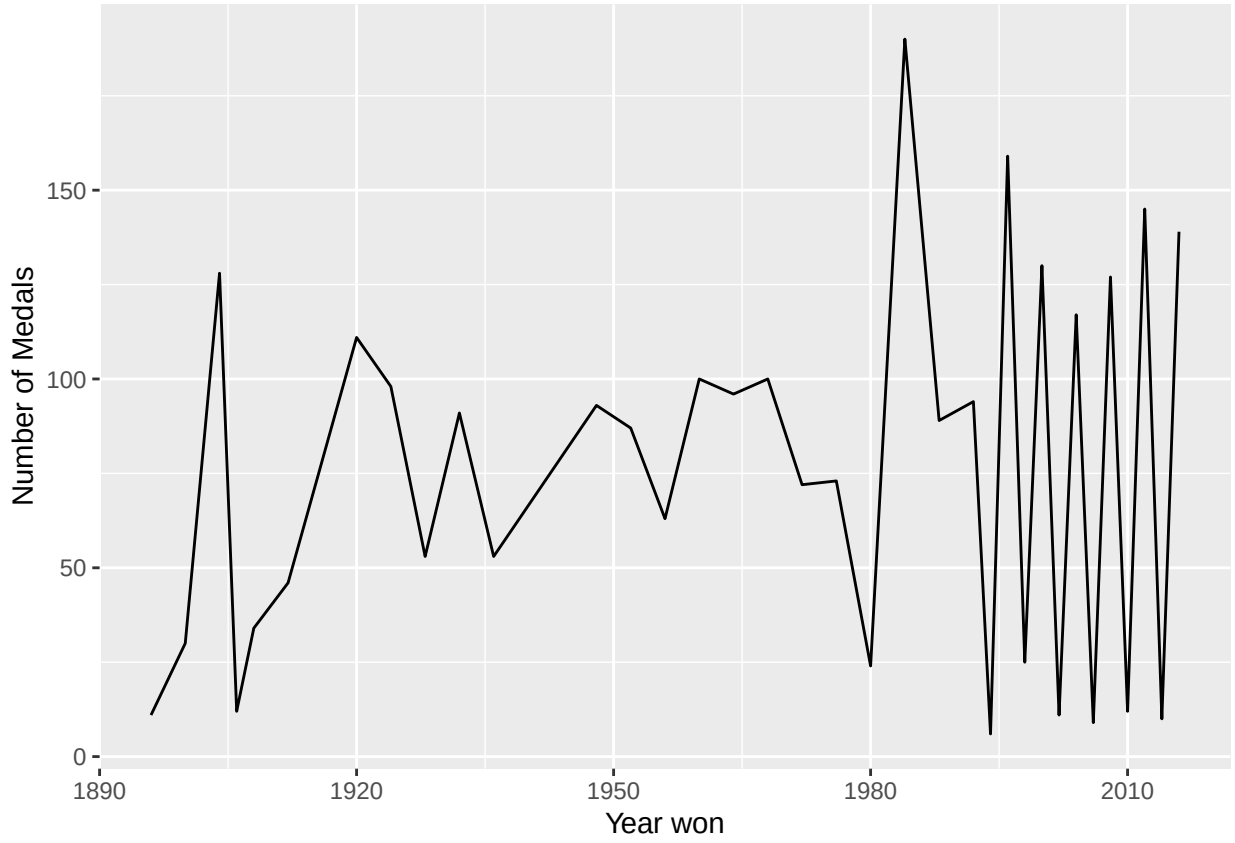
the most appropriate graph would be a scatter plot. In order to correct for athletes with the same age we could use both the jitter function and changing the transparency of the function. With this data I think changing the transparency makes more sense

Question 3:

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

```
## # A tibble: 35 x 2
##     year num_medals
##    <dbl>      <int>
##  1  1896         11
##  2  1900         30
##  3  1904        128
##  4  1906         12
##  5  1908         34
##  6  1912         46
##  7  1920        111
##  8  1924         98
##  9  1928         53
## 10  1932         91
## # ... with 25 more rows
```

```
ggplot(data = us_medals,
       mapping = aes(x = year, y = num_medals)) +
  labs(y = "Number of Medals", x = "Year won") +
  geom_line()
```




```
us_medals$year[us_medals$num_medals == max(us_medals$num_medals)]
```

```
## [1] 1984
```

```
    ##okay so i am looking for what year the max number of medals for the US
```

The countries most successful year would be 1984. You can visualize this via the graph or just find the Max using line of code above in line 124. I bet its because that we are not good at winter sports.

Question 4:

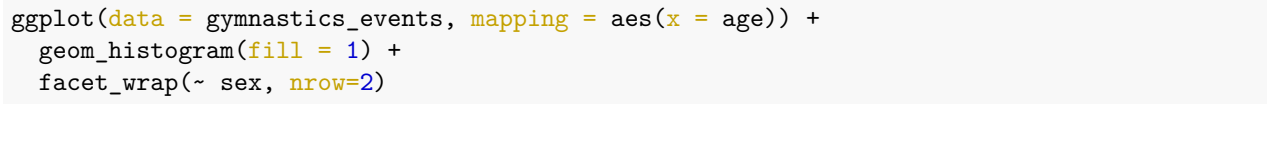
```
two_events <- gold_medalist %>%
  filter(
    event == "Gymnastics Men's Individual All-Around"  |
      event == "Gymnastics Women's Individual All-Around" |
      event == "Athletics Women's 100 metres"    |
      event == "Athletics Men's 100 metres"
  )
glimpse(two_events)
```

```
## Rows: 94
## Columns: 15
## $ id     <dbl> 519, 3281, 4198, 5396, 5547, 6890, 11495, 12166, 13029, 13029, ~
## $ name   <chr> "Harold Maurice Abrahams", "Simona Amnar (-Tabr)", "Nikolay Yef~
## $ sex    <chr> "M", "F", "M", "M", "F", "M", "F", "F", "M", "M", "M", "M", "M"~
## $ age    <dbl> 24, 20, 23, 23, 27, 28, 19, 30, 21, 25, 29, 22, 25, 29, 21, 22,~
## $ height <dbl> 183, 158, 166, 167, 165, 183, 143, 175, 196, 196, 196, 183, 167~
## $ weight <dbl> 75, 44, 60, 63, 52, 82, 47, 63, 95, 95, 95, 80, NA, NA, 66, 58,~
## $ team   <chr> "Great Britain", "Romania", "Soviet Union", "Soviet Union", "Un~
## $ noc    <chr> "GBR", "ROU", "URS", "URS", "USA", "CAN", "USA", "NED", "JAM", ~
## $ games  <chr> "1924 Summer", "2000 Summer", "1976 Summer", "1988 Summer", "19~
## $ year   <dbl> 1924, 2000, 1976, 1988, 1984, 1996, 2016, 1948, 2008, 2012, 201~
## $ season <chr> "Summer", "Summer", "Summer", "Summer", "Summer", "Summer", "Su~
## $ city   <chr> "Paris", "Sydney", "Montreal", "Seoul", "Los Angeles", "Atlanta~
## $ sport  <chr> "Athletics", "Gymnastics", "Gymnastics", "Gymnastics", "Athleti~
## $ event  <chr> "Athletics Men's 100 metres", "Gymnastics Women's Individual Al~
## $ medal  <chr> "Gold", "Gold", "Gold", "Gold", "Gold", "Gold", "Gold", "Gold",~
```

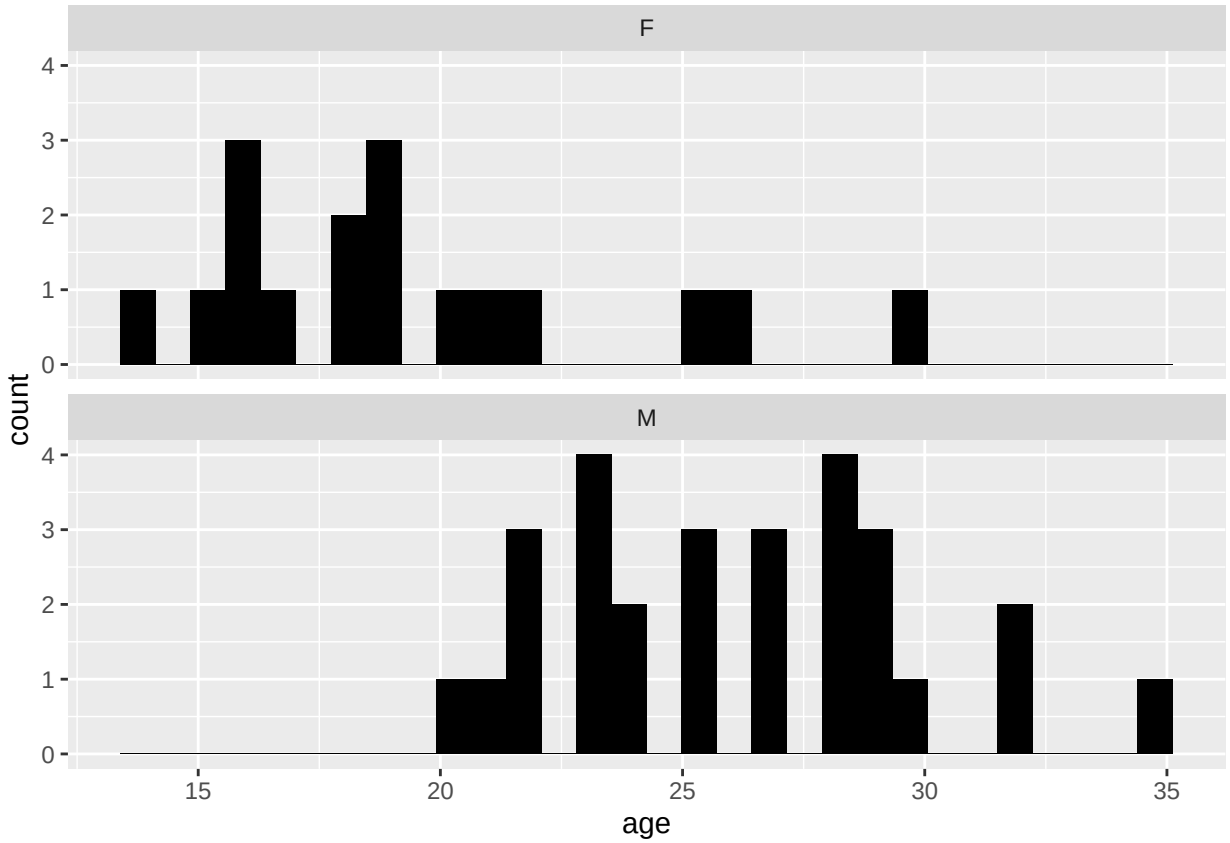
```
gymnastics_events <- two_events %>%
  filter(
    event == "Gymnastics Men's Individual All-Around"  |
      event == "Gymnastics Women's Individual All-Around"
  )
glimpse(gymnastics_events)
```

```
## Rows: 45
## Columns: 15
## $ id     <dbl> 3281, 4198, 5396, 11495, 14549, 14549, 18826, 18826, 21402, 214~
## $ name   <chr> "Simona Amnar (-Tabr)", "Nikolay Yefimovich Andrianov", "Vladim~
## $ sex    <chr> "F", "M", "M", "F", "M", "M", "F", "F", "M", "M", "F", "F", "M"~
## $ age    <dbl> 20, 23, 23, 19, 25, 29, 22, 26, 30, 35, 14, 18, 22, 16, 27, 27,~
## $ height <dbl> 158, 166, 167, 143, 167, 167, 160, 160, NA, NA, 162, 148, 178, ~
## $ weight <dbl> 44, 60, 63, 47, NA, NA, 58, 58, NA, NA, 45, 45, 70, 50, 58, 58,~
## $ team   <chr> "Romania", "Soviet Union", "Soviet Union", "United States", "It~
## $ noc    <chr> "ROU", "URS", "URS", "USA", "ITA", "ITA", "TCH", "TCH", "URS", ~
## $ games  <chr> "2000 Summer", "1976 Summer", "1988 Summer", "2016 Summer", "19~
## $ year   <dbl> 2000, 1976, 1988, 2016, 1908, 1912, 1964, 1968, 1952, 1956, 197~
## $ season <chr> "Summer", "Summer", "Summer", "Summer", "Summer", "Summer", "Su~
## $ city   <chr> "Sydney", "Montreal", "Seoul", "Rio de Janeiro", "London", "Sto~
## $ sport  <chr> "Gymnastics", "Gymnastics", "Gymnastics", "Gymnastics", "Gymnas~
## $ event  <chr> "Gymnastics Women's Individual All-Around", "Gymnastics Men's I~
## $ medal  <chr> "Gold", "Gold", "Gold", "Gold", "Gold", "Gold", "Gold", "Gold",~
```

```
ggplot(data = gymnastics_events, mapping = aes(x = age)) +
  geom_histogram(fill = 1) +
  facet_wrap(~ sex, nrow=2)
```



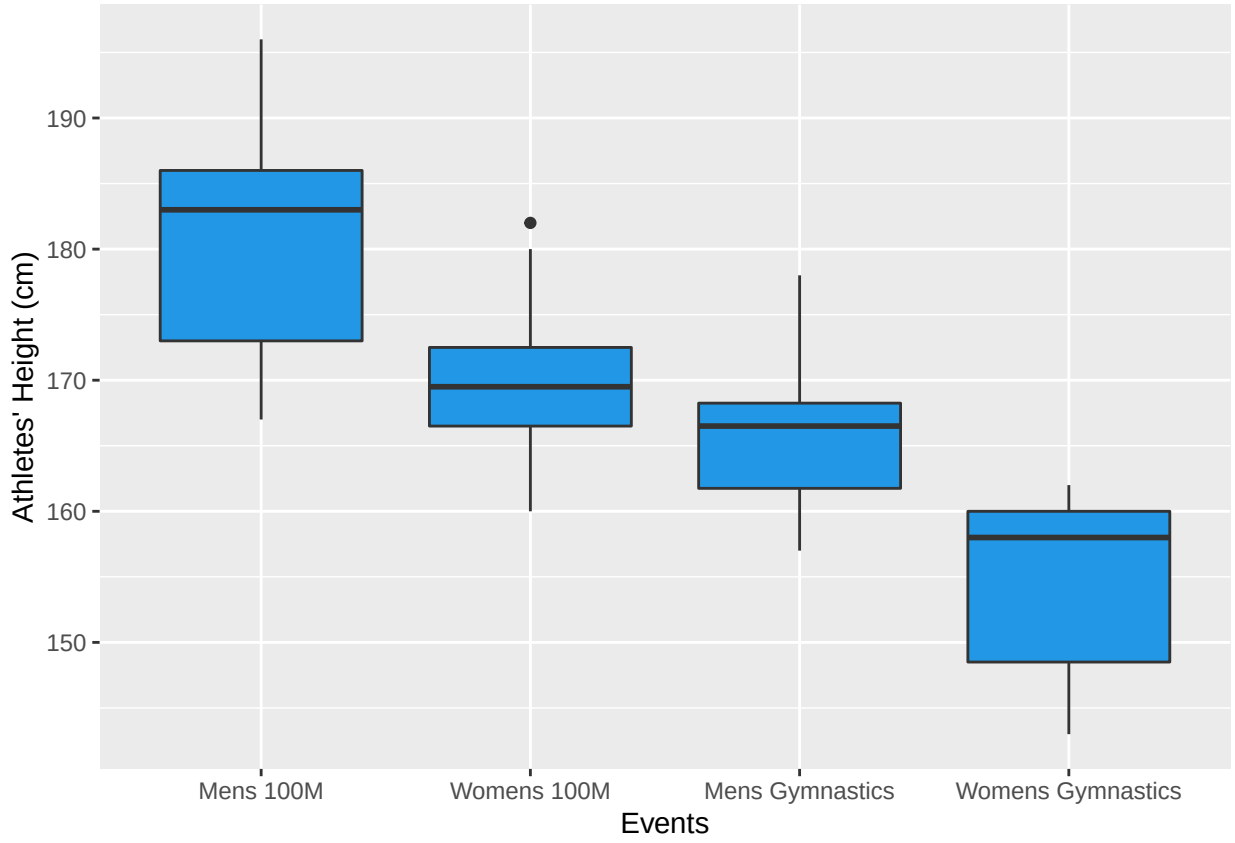
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Question 5:

```
ggplot(data = two_events, mapping = aes(y = height, x = event)) +
  labs(x = "Events", y = "Athletes' Height (cm)") +
  scale_x_discrete(labels = c('Mens 100M','Womens 100M','Mens Gymnastics', 'Womens Gymnastics')) +
  geom_boxplot(fill = 4)
```

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




```
?geom_boxplot
```

The mens 100 meter dash has the tallest athletes.

Question 6:

```
us_medalists <- gold_medalist %>%
  filter(noc == "USA")
ggplot(data = us_medalists, aes(x = year, fill = sex)) +
  labs(y= "Number of Gold Medalists", x = "Year") +
  geom_bar(position = "dodge")
```

That there were no women in the olmpics in the early years of dawn.