Class19Miniproject

R Info

```
library("datapasta")
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
```

Data scraping

```
cdc <- data.frame(
                                    Year = c(1922L,1923L,1924L,1925L,
                                             1926L,1927L,1928L,1929L,1930L,1931L,
                                             1932L,1933L,1934L,1935L,1936L,
                                             1937L,1938L,1939L,1940L,1941L,1942L,
                                             1943L,1944L,1945L,1946L,1947L,
                                             1948L,1949L,1950L,1951L,1952L,
                                             1953L,1954L,1955L,1956L,1957L,1958L,
                                             1959L,1960L,1961L,1962L,1963L,
                                             1964L,1965L,1966L,1967L,1968L,1969L,
```

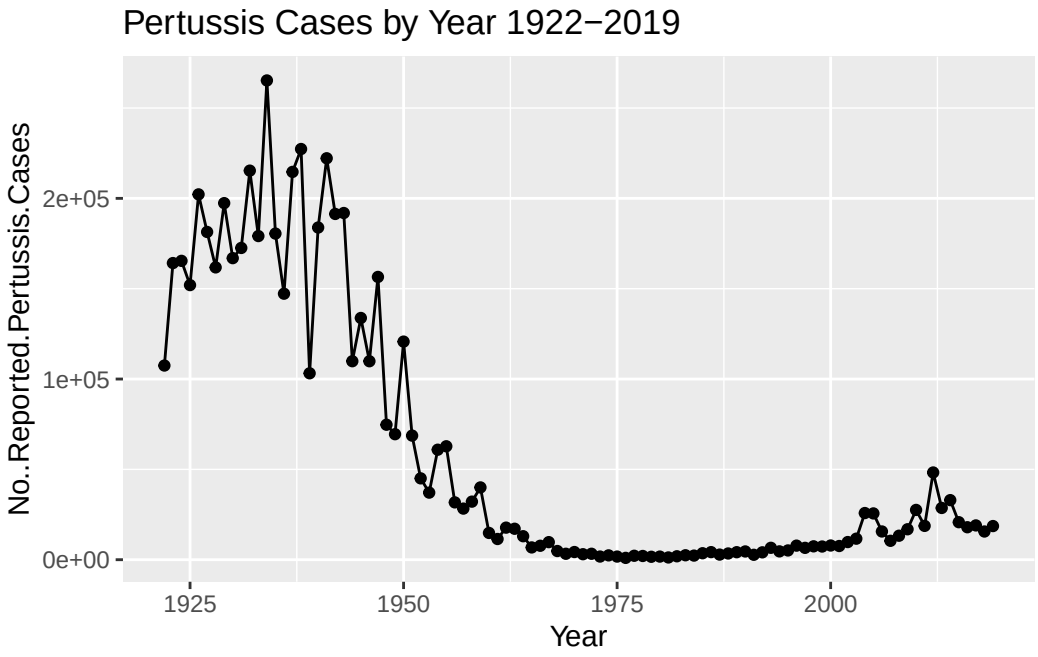
```
                                        1970L,1971L,1972L,1973L,1974L,
                                        1975L,1976L,1977L,1978L,1979L,1980L,
                                        1981L,1982L,1983L,1984L,1985L,
                                        1986L,1987L,1988L,1989L,1990L,
                                        1991L,1992L,1993L,1994L,1995L,1996L,
                                        1997L,1998L,1999L,2000L,2001L,
                                        2002L,2003L,2004L,2005L,2006L,2007L,
                                        2008L,2009L,2010L,2011L,2012L,
                                        2013L,2014L,2015L,2016L,2017L,2018L,
                                        2019L),
        No..Reported.Pertussis.Cases = c(107473,164191,165418,152003,
                                        202210,181411,161799,197371,
                                        166914,172559,215343,179135,265269,
                                        180518,147237,214652,227319,103188,
                                        183866,222202,191383,191890,109873,
                                        133792,109860,156517,74715,69479,
                                        120718,68687,45030,37129,60886,
                                        62786,31732,28295,32148,40005,
                                        14809,11468,17749,17135,13005,6799,
                                        7717,9718,4810,3285,4249,3036,
                                        3287,1759,2402,1738,1010,2177,2063,
                                        1623,1730,1248,1895,2463,2276,
                                        3589,4195,2823,3450,4157,4570,
                                        2719,4083,6586,4617,5137,7796,6564,
                                        7405,7298,7867,7580,9771,11647,
                                        25827,25616,15632,10454,13278,
                                        16858,27550,18719,48277,28639,32971,
                                        20762,17972,18975,15609,18617)
       )
```

ploting

Q1

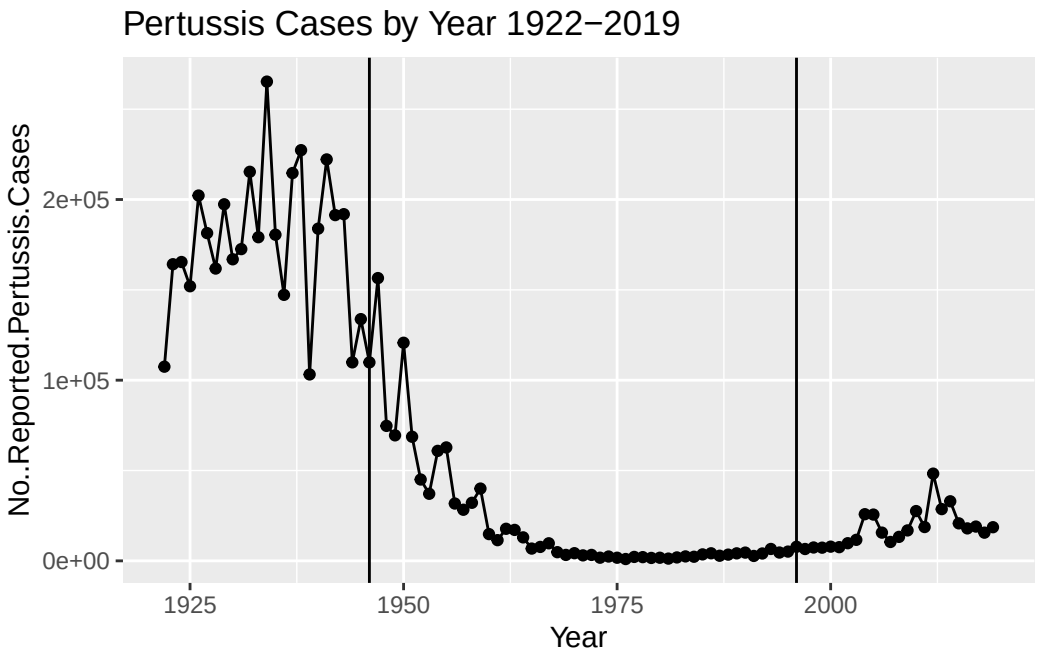
```
library("ggplot2")

casebyyear <- ggplot(cdc) + aes(Year,No..Reported.Pertussis.Cases)+
  geom_point()+
  geom_line()+
```

```
  labs(title="Pertussis Cases by Year 1922-2019",xlab="Year",ylab="Number of cases")
casebyyear
```



Q2

```
casebyyear+(geom_vline(xintercept=c(1946,1996)))
```



Q3 What happened after introduction of the Ap vaccine?

We see an increase in number of cases. There was a celar decline after the introduction of the WP vaccine, and an increase after the switch to the aP vaccine

Q3

```
library("jsonlite")

subject <- read_json("https://www.cmi-pb.org/api/subject", simplifyVector = TRUE)

head(subject, 3)
```

```
  subject_id infancy_vac biological_sex              ethnicity  race
1          1          wP         Female Not Hispanic or Latino White
2          2          wP         Female Not Hispanic or Latino White
3          3          wP         Female                 Unknown White
  year_of_birth date_of_boost      dataset
1    1986-01-01    2016-09-12 2020_dataset
```

```
2    1968-01-01    2019-01-28 2020_dataset
3    1983-01-01    2016-10-10 2020_dataset
```

Q4 How many aP and wP infancy vaccinated subjects are in the dataset?

```
table(subject$infancy_vac)
```

```
aP wP
47 49
```

Q5. How many Male and Female subjects/patients are in the dataset?

```
table(subject$biological_sex)
```

```
Female   Male
    66     30
```

Q6. What is the breakdown of race and biological sex (e.g. number of Asian females, White males etc...)?

```
table(subject$ethnicity)
```

```
   Hispanic or Latino Not Hispanic or Latino                Unknown
                   23                      69                      4
```

```
table(subject$race)
```

```
             American Indian/Alaska Native
                                         1
                                     Asian
                                        27
                 Black or African American
                                         2
                           More Than One Race
                                        10
Native Hawaiian or Other Pacific Islander
                                         2
                    Unknown or Not Reported
                                        14
                                     White
                                        40
```

```
library("lubridate")
```

```
Loading required package: timechange
```

```
Attaching package: 'lubridate'
```

```
The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

Q7 Using this approach determine (i) the average age of wP individuals, (ii) the average age of aP individuals; and (iii) are they significantly different?

```
subject$age <- time_length(today()-ymd(subject$year_of_birth),'years')

subject$age_of_boost <- time_length(ymd(subject$date_of_boost)-ymd(subject$year_of_birth),
wp_ind <- subject %>% filter(infancy_vac=="wP")
ap_ind <- subject %>% filter(infancy_vac=="aP")

mean(wp_ind$age)
```

```
[1] 36.07532
```

```
mean(ap_ind$age)
```

```
[1] 25.23087
```

```
t.test(wp_ind$age,ap_ind$age)
```

```
    Welch Two Sample t-test

data:  wp_ind$age and ap_ind$age
t = 12.092, df = 51.082, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  9.044045 12.644857
sample estimates:
mean of x mean of y
 36.07532  25.23087
```

Q8. Determine the age of all individuals at time of boost?

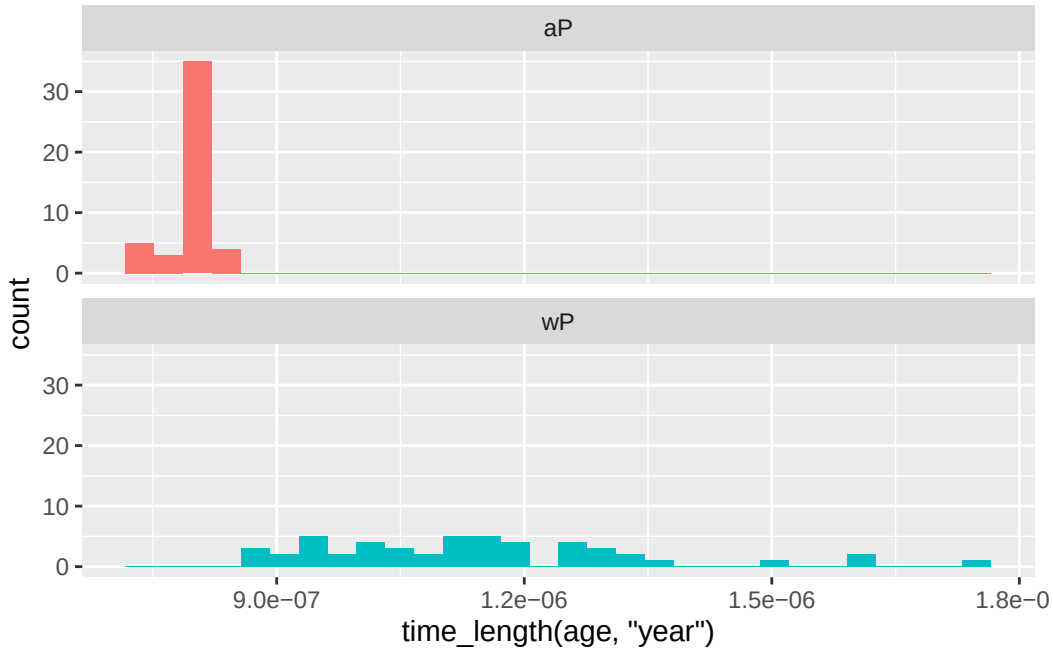
```
head(subject$age_of_boost)
```

```
[1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481
```

Q9. With the help of a faceted boxplot or histogram (see below), do you think these two groups are significantly different?

```
ggplot(subject) +
  aes(time_length(age, "year"),
      fill=as.factor(infancy_vac)) +
  geom_histogram(show.legend=FALSE) +
  facet_wrap(vars(infancy_vac), nrow=2)
```

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



Dyplr function of *join

`inner_join()` joins everything that is the same row `full_join()` this allows two datasets to join and keep everything

We want to "join" the `subject` and `specimin` tables to have all the meta data we need tofr later analysis. We can use the dyplr function `*_join` function for this task.

Q10

```
specimen <- read_json("https://www.cmi-pb.org/api/specimen", simplifyVector = TRUE)


meta <- inner_join(specimen, subject)
```

```
Joining, by = "subject_id"
```

```
titer <- read_json("https://www.cmi-pb.org/api/ab_titer", simplifyVector = TRUE)
head(titer)
```

```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1     IgE               FALSE   Total 1110.21154       2.493425
2           1     IgE               FALSE   Total 2708.91616       2.493425
3           1     IgG                TRUE      PT   68.56614       3.736992
4           1     IgG                TRUE     PRN  332.12718       2.602350
5           1     IgG                TRUE     FHA 1887.12263      34.050956
6           1     IgE                TRUE     ACT    0.10000       1.000000
   unit lower_limit_of_detection
1 UG/ML                 2.096133
2 IU/ML                29.170000
3 IU/ML                 0.530000
4 IU/ML                 6.205949
5 IU/ML                 4.679535
6 IU/ML                 2.816431
```

```
abdata <- inner_join(meta,titer)
```

```
Joining, by = "specimen_id"
```

Q11. How many specimens (i.e. entries in abdata) do we have for each isotype?

```
table(abdata$isotype)
```

```
 IgE  IgG IgG1 IgG2 IgG3 IgG4
6698 1413 6141 6141 6141 6141
```

Q12. What do you notice about the number of visit 8 specimens compared to other visits?

```
table(abdata$visit)
```

```
   1    2    3    4    5    6    7    8
5795 4640 4640 4640 4640 4320 3920   80
```

Its much smaller!

```
ig1 <- abdata %>% filter(isotype == "IgG1", visit!=8)
head(ig1)
```

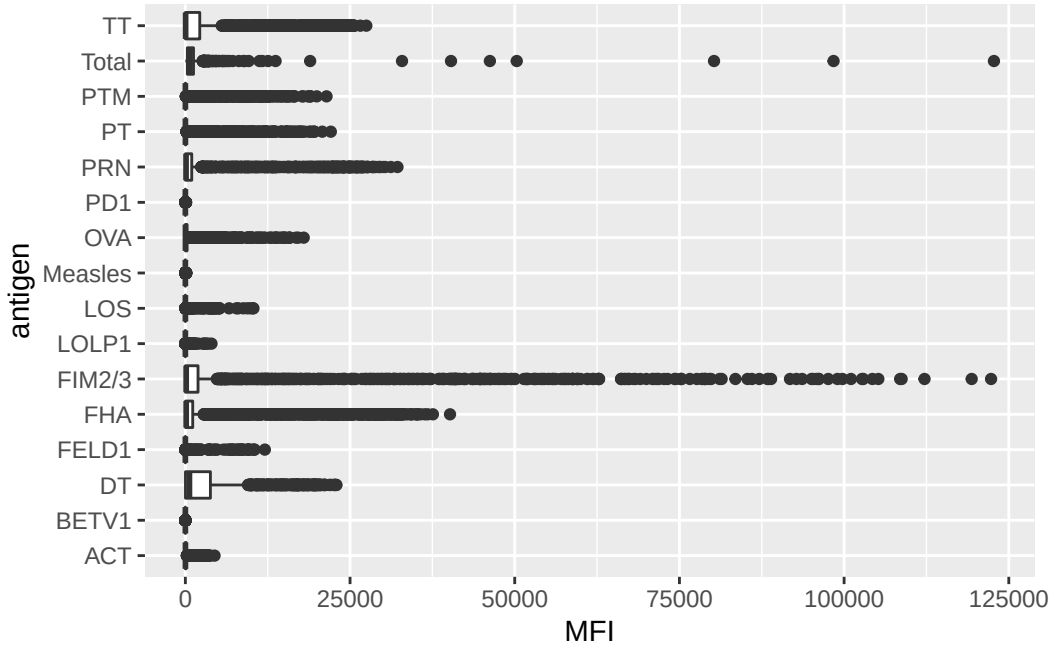
```
  specimen_id subject_id actual_day_relative_to_boost
1           1          1                           -3
2           1          1                           -3
3           1          1                           -3
4           1          1                           -3
5           1          1                           -3
6           1          1                           -3
  planned_day_relative_to_boost specimen_type visit infancy_vac biological_sex
1                             0         Blood     1          wP         Female
2                             0         Blood     1          wP         Female
3                             0         Blood     1          wP         Female
4                             0         Blood     1          wP         Female
5                             0         Blood     1          wP         Female
6                             0         Blood     1          wP         Female
               ethnicity  race year_of_birth date_of_boost      dataset     age
1 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.9117
2 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.9117
3 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.9117
4 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.9117
5 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.9117
6 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.9117
  age_of_boost isotype is_antigen_specific antigen        MFI MFI_normalised
1     30.69678    IgG1                TRUE     ACT 274.355068      0.6928058
2     30.69678    IgG1                TRUE     LOS  10.974026      2.1645083
3     30.69678    IgG1                TRUE   FELD1   1.448796      0.8080941
4     30.69678    IgG1                TRUE   BETV1   0.100000      1.0000000
```

```
5     30.69678    IgG1                TRUE   LOLP1   0.100000      1.0000000
6     30.69678    IgG1                TRUE Measles  36.277417      1.6638332
   unit lower_limit_of_detection
1 IU/ML                 3.848750
2 IU/ML                 4.357917
3 IU/ML                 2.699944
4 IU/ML                 1.734784
5 IU/ML                 2.550606
6 IU/ML                 4.438966
```

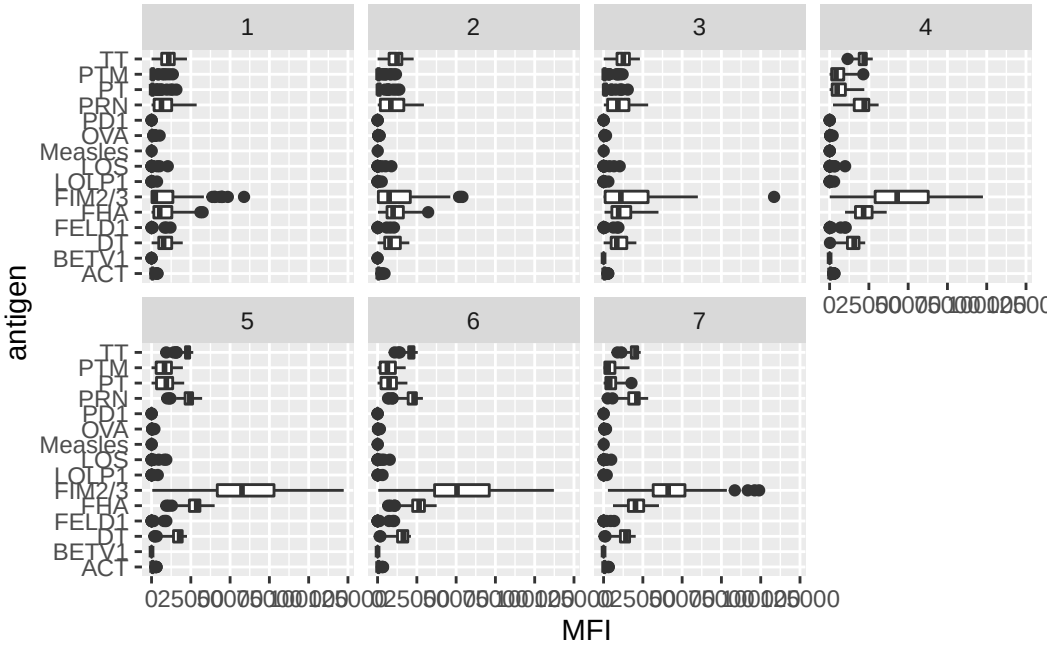
Q13

```
abdataboxplot <- ggplot(abdata)+
  aes(MFI, antigen) + geom_boxplot()
abdataboxplot
```

```
Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```



```
abdataboxplot_norm <- ggplot(abdata)+
  aes(MFI_normalised, antigen) + geom_boxplot()


ggplot(ig1)+aes(MFI,antigen) +geom_boxplot()+facet_wrap(vars(visit),nrow=2)
```




```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.

rna <- read_json(url, simplifyVector = TRUE)
```

```
ssrna <- inner_join(rna, meta)
```

```
Joining, by = "specimen_id"
```