# class 19 Investigating Pertussis Resurgence

## 1. Investigating pertussis cases by year

Use the `datapasta` package to get the data from CDC Website

> Q1. With the help of the R "addin" package datapasta assign the CDC pertussis case number data to a data frame called `cdc` and use ggplot to make a plot of cases numbers over time.

```
  Year No..Reported.Pertussis.Cases
1 1922                       107473
2 1923                       164191
3 1924                       165418
4 1925                       152003
5 1926                       202210
6 1927                       181411
```

Make the ggplot

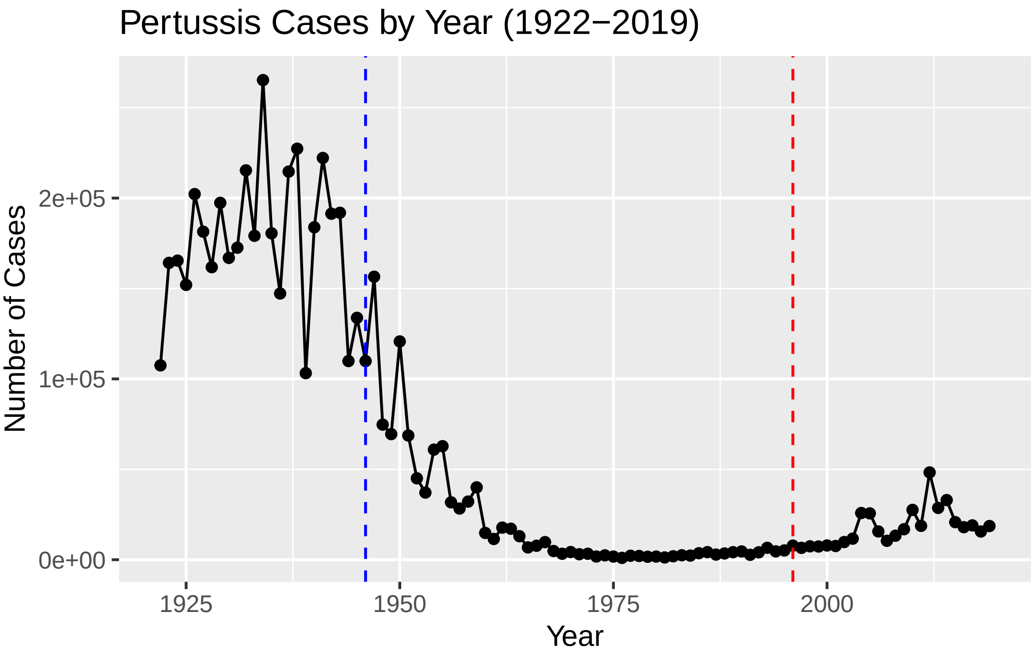
```r
library(ggplot2)
```

```r
basePertussisPlot <- ggplot(cdc) +
  aes(Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +
  labs(x = "Year", y = "Number of Cases", title = "Pertussis Cases by Year (1922-2019)")
```

## 2. A Tale of Two Vaccines (wP & aP)

> Q2. Using the ggplot geom_vline() function add lines to your previous plot for the 1946 introduction of the wP vaccine and the 1996 switch to aP vaccine (see example in the hint below). What do you notice?

```
basePertussisPlot +
  geom_vline(xintercept = 1946, linetype = "dashed", col = "blue") +
  geom_vline(xintercept = 1996, linetype = "dashed", col = "red")
```



> Q3. Describe what happened after the introduction of the aP vaccine? Do you have a possible explanation for the observed trend?

After the introduction of the aP vaccine, the cases for pertussis has started rising again with a peak around 2012 which was the highest since the mid 1960s. This could be from multiple reasons such as hesitancy to vaccinate, bacterial immunity from the vaccine, different variants of the bacteria, etc.

## 3. Exploring CMI-PB Data

The CMB-PB project provides long term data about the pertussis resurgence.

## The CMI-PB API returns JSON data

```
# Allows us to read, write and process JSON data
library(jsonlite)
```

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

head(subject)
```

```
  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
4          4          wP           Male Not Hispanic or Latino Asian
5          5          wP           Male Not Hispanic or Latino Asian
6          6          wP         Female Not Hispanic or Latino 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
4    1988-01-01    2016-08-29 2020_dataset
5    1991-01-01    2016-08-29 2020_dataset
6    1988-01-01    2016-10-10 2020_dataset
```

> Q4. How may 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$race, subject$biological_sex)
```

```
                                            Female Male
  American Indian/Alaska Native                  0    1
  Asian                                         18    9
  Black or African American                      2    0
  More Than One Race                             8    2
  Native Hawaiian or Other Pacific Islander      1    1
  Unknown or Not Reported                       10    4
  White                                         27   13
```

## Side-Note: Working with dates

```
library(lubridate)
```

```
Loading required package: timechange


Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

```
today()
```

```
[1] "2022-11-29"
```

```
today() - ymd("2000-01-01")
```

```
Time difference of 8368 days
```

```
time_length ( today() - ymd("2000-01-01"), "years" )
```

```
[1] 22.91034
```

> 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?

```
# Use todays date to calculate age in days
subject$age <- today() - ymd(subject$year_of_birth)

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
```

```
# aP data

ap <- subject %>% filter(infancy_vac == "aP")

round( summary( time_length( ap$age, "years" ) ) )
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     23      25      26      25      26      27 
```

```
# wP data

wp <- subject %>% filter(infancy_vac == "wP")

round ( summary( time_length( wp$age, "years") ) )
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     28      32      35      36      40      55 
```

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

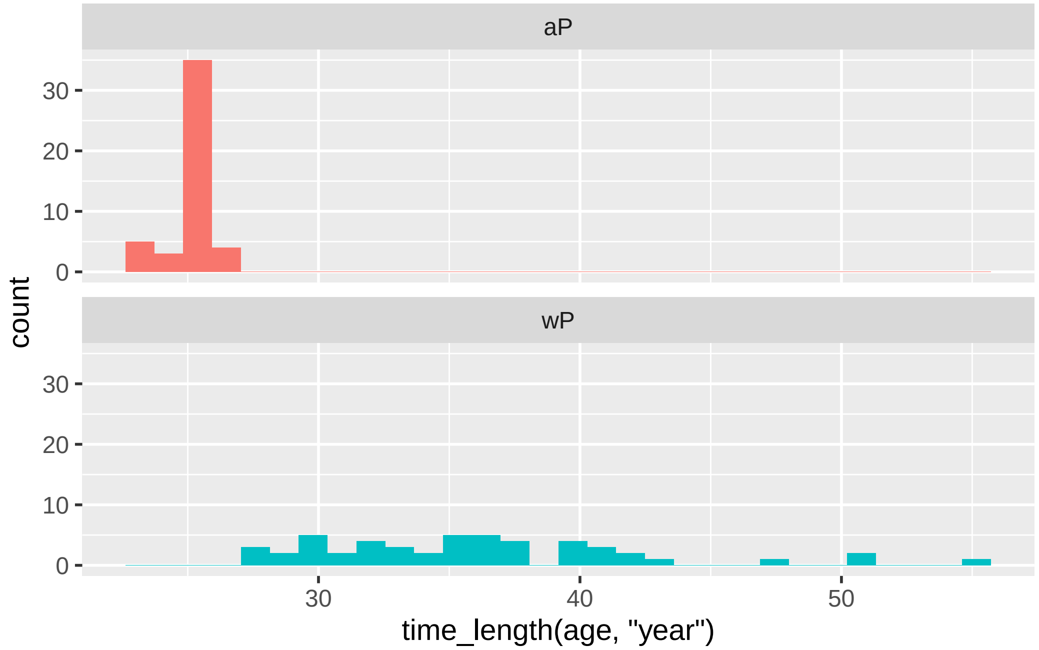
```r
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
round(age_at_boost)
```

```
 [1] 31 51 34 29 26 29 36 34 21 35 31 35 20 24 28 30 37 20 23 32 26 24 26 29 43
[26] 47 47 29 21 21 28 24 24 21 21 31 26 32 27 26 21 20 22 19 21 19 19 22 20 21
[51] 19 23 20 21 19 36 34 32 26 25 29 34 20 35 20 29 28 20 27 34 26 20 19 20 32
[76] 23 32 20 19 19 20 19 21 19 20 20 20 19 19 20 20 20 21 20 20 20
```

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

```r
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`.
```



```r
# Calculate p-value
x <- t.test(time_length( wp$age, "years" ),
       time_length( ap$age, "years" ))

x$p.value
```

```
[1] 1.316045e-16
```

## Joining Multiple Tables

```r
# Complete the API URLs...
specimen <- read_json("http://cmi-pb.org/api/specimen", simplifyVector = TRUE)
titer <- read_json("http://cmi-pb.org/api/ab_titer", simplifyVector = TRUE)
```

Use the `dplyr` package in order to join the tables together.

> Q9. Complete the code to join `specimen` and `subject` tables to make a new merged data frame containing all specimen records along with their associated subject details:

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

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

```
dim(meta)
```

```
[1] 729  14
```

```
head(meta)
```

```
  specimen_id subject_id actual_day_relative_to_boost
1           1          1                           -3
2           2          1                          736
3           3          1                            1
4           4          1                            3
5           5          1                            7
6           6          1                           11
  planned_day_relative_to_boost specimen_type visit infancy_vac biological_sex
1                             0         Blood     1          wP         Female
2                           736         Blood    10          wP         Female
3                             1         Blood     2          wP         Female
4                             3         Blood     3          wP         Female
5                             7         Blood     4          wP         Female
6                            14         Blood     5          wP         Female
               ethnicity  race year_of_birth date_of_boost      dataset
1 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
2 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
3 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
4 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
5 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
6 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
         age
1 13481 days
2 13481 days
3 13481 days
4 13481 days
5 13481 days
6 13481 days
```

> Q10. Now using the same procedure join meta with titer data so we can further analyze this data in terms of time of visit aP/wP, male/female etc.

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

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

```r
dim(abdata)
```

```
[1] 32675    21
```

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

```r
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?

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

```

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

There are a lot less visit 8 specimens compared to the other visit specimens.

## 4. Examine IgG1 Ab titer levels

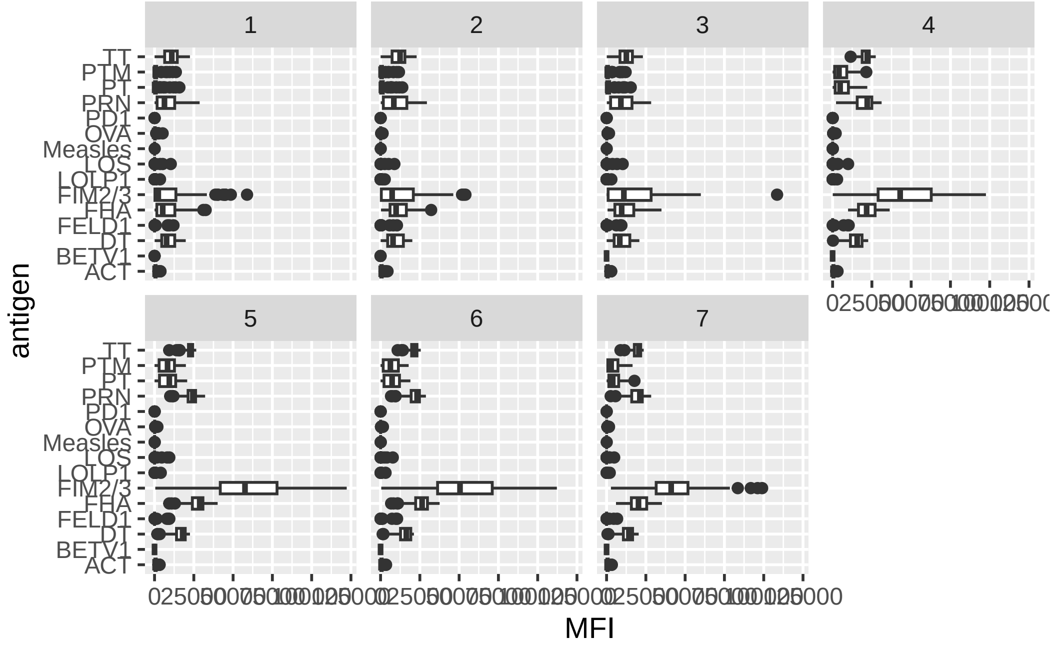
Now using our joined/merged/linked abdata dataset filter() for IgG1 isotype and exclude the small number of visit 8 entries.

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

```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1    IgG1                TRUE     ACT 274.355068      0.6928058
2           1    IgG1                TRUE     LOS  10.974026      2.1645083
3           1    IgG1                TRUE   FELD1   1.448796      0.8080941
4           1    IgG1                TRUE   BETV1   0.100000      1.0000000
5           1    IgG1                TRUE   LOLP1   0.100000      1.0000000
6           1    IgG1                TRUE Measles  36.277417      1.6638332
   unit lower_limit_of_detection subject_id actual_day_relative_to_boost
1 IU/ML                 3.848750          1                           -3
2 IU/ML                 4.357917          1                           -3
3 IU/ML                 2.699944          1                           -3
4 IU/ML                 1.734784          1                           -3
5 IU/ML                 2.550606          1                           -3
6 IU/ML                 4.438966          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
1 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
2 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
3 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
4 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
5 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
6 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
         age
1 13481 days
2 13481 days
3 13481 days
4 13481 days
5 13481 days
6 13481 days
```

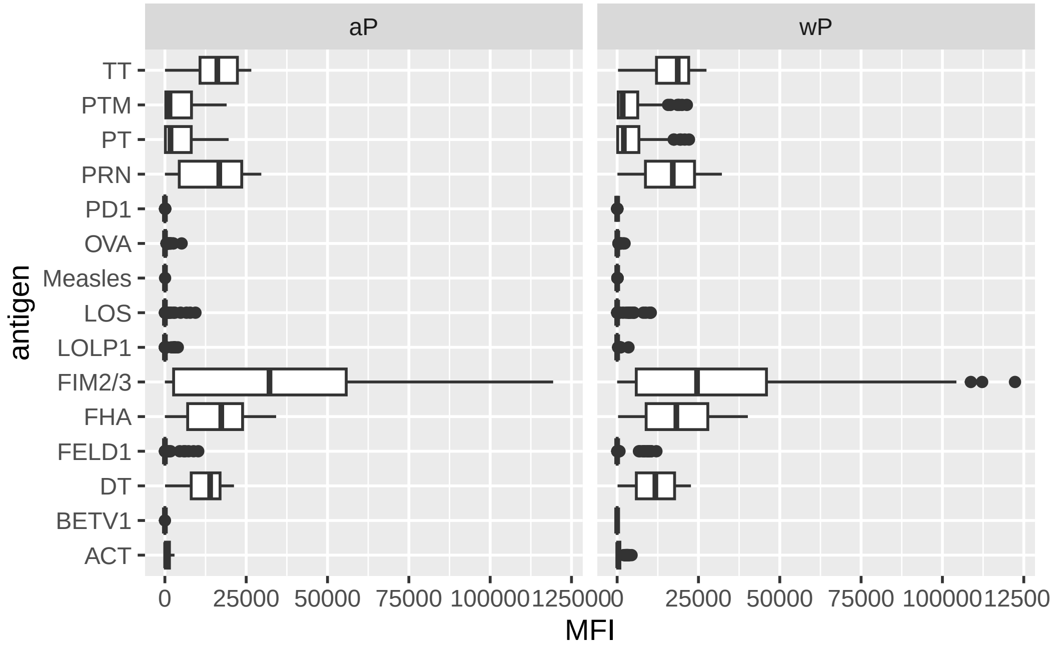
Q13. Complete the following code to make a summary boxplot of Ab titer levels for all antigens:

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



Now facet by aP and wP.

```
ggplot(ig1) +
  aes(MFI, antigen) +
  geom_boxplot() +
  facet_wrap(vars(infancy_vac))
```



Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

The FIM2/3 antigens show differences in the level of IgG1 antibody titers recognizing them.