Class18: Investigating Pertussis Resurgence

Richard Gao (PID: A16490010)

First we will examine and explore Pertussis cas numbers in the US as tracked by the CDC: https://www.cdc.gov/pertussis/surv-reporting/cases-by-year.html

We can use the datapastapackage to scrape this data from the website into R:

```
cdc <- data.frame( year = c(1922L,1923L,1924L,1925L, 1926L,1927L,1928L,1929L,1930L,1931L,1
  cases = c(107473,164191,165418,152003,202210,181411,161799,197371,166914,172559,215343,1
       )
```

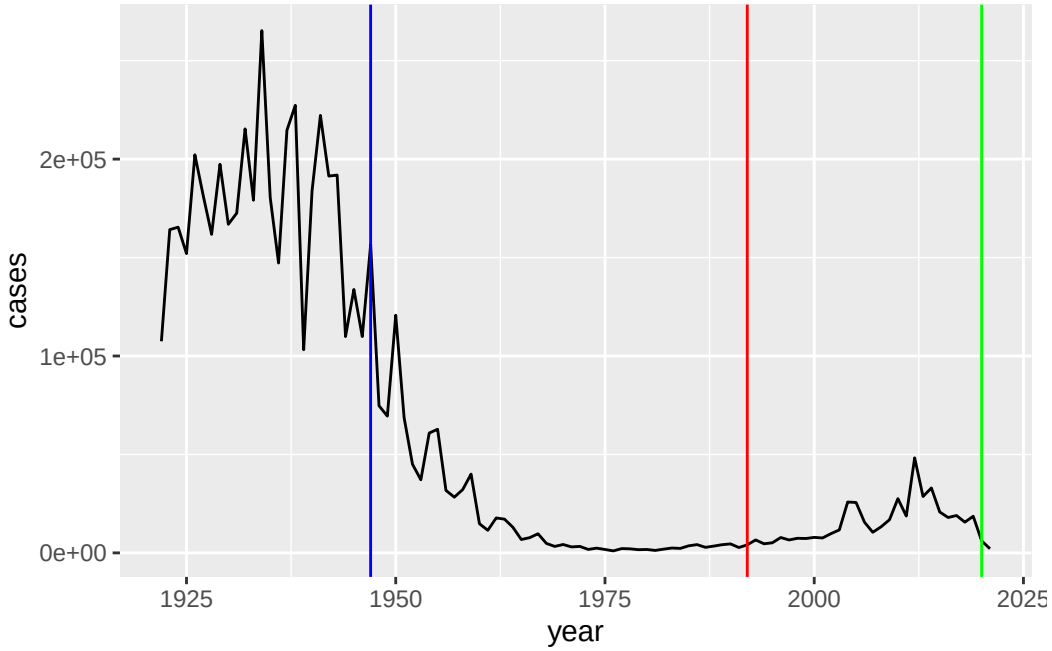
```
head(cdc)
```

```
  year  cases
1 1922 107473
2 1923 164191
3 1924 165418
4 1925 152003
5 1926 202210
6 1927 181411
```

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.

```
library(ggplot2)
```

```
ggplot(cdc, aes(year, cases)) +
  geom_line() +
  geom_vline(xintercept=1947, col="blue") +
  geom_vline(xintercept=1992, col="red") +
  geom_vline(xintercept=2020, col="green")
```



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?

I notice that between the 1946 introduction of the wP vaccine and the 1996 switch to the aP vaccine, the number of pertussis cases dropped significantly and stayed low.

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 number of perstussis cases resurged. A possible explanation for this is that the immunity gained from the aP vaccine doesn't last very long compared to the original wP vaccine.

Access data from the CMI-PB project

This database (like many modern projects) uses an APi to resturn JSON format data.

We will use the R package `jsonlite`.

```
library(jsonlite)
subject <- read_json("http://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 many aP and wP infancy vaccinated subjects are in the dataset?

```
sum(subject$infancy_vac == "wP")
```

```
[1] 58
```

```
sum(subject$infancy_vac == "aP")
```

```
[1] 60
```

or just do:

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

```
aP wP
60 58
```

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

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

```
Female   Male
    79     39
```

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                                       21   11
Black or African American                    2    0
More Than One Race                           9    2
Native Hawaiian or Other Pacific Islander    1    1
Unknown or Not Reported                     11    4
White                                       35   20
```

```
subject$year_of_birth
```

```
  [1] "1986-01-01" "1968-01-01" "1983-01-01" "1988-01-01" "1991-01-01"
  [6] "1988-01-01" "1981-01-01" "1985-01-01" "1996-01-01" "1982-01-01"
 [11] "1986-01-01" "1982-01-01" "1997-01-01" "1993-01-01" "1989-01-01"
 [16] "1987-01-01" "1980-01-01" "1997-01-01" "1994-01-01" "1981-01-01"
 [21] "1983-01-01" "1985-01-01" "1991-01-01" "1992-01-01" "1988-01-01"
 [26] "1983-01-01" "1997-01-01" "1982-01-01" "1997-01-01" "1988-01-01"
 [31] "1989-01-01" "1997-01-01" "1990-01-01" "1983-01-01" "1991-01-01"
 [36] "1997-01-01" "1998-01-01" "1997-01-01" "1985-01-01" "1994-01-01"
 [41] "1985-01-01" "1997-01-01" "1998-01-01" "1998-01-01" "1997-01-01"
 [46] "1998-01-01" "1996-01-01" "1998-01-01" "1997-01-01" "1997-01-01"
 [51] "1997-01-01" "1998-01-01" "1998-01-01" "1997-01-01" "1997-01-01"
 [56] "1997-01-01" "1996-01-01" "1997-01-01" "1997-01-01" "1997-01-01"
 [61] "1987-01-01" "1993-01-01" "1995-01-01" "1993-01-01" "1990-01-01"
 [66] "1976-01-01" "1972-01-01" "1972-01-01" "1990-01-01" "1998-01-01"
 [71] "1998-01-01" "1991-01-01" "1995-01-01" "1995-01-01" "1998-01-01"
 [76] "1998-01-01" "1988-01-01" "1993-01-01" "1987-01-01" "1992-01-01"
 [81] "1993-01-01" "1998-01-01" "1999-01-01" "1997-01-01" "2000-01-01"
 [86] "1998-01-01" "2000-01-01" "2000-01-01" "1997-01-01" "1999-01-01"
 [91] "1998-01-01" "2000-01-01" "1996-01-01" "1999-01-01" "1998-01-01"
 [96] "2000-01-01" "1986-01-01" "1993-01-01" "1999-01-01" "2001-01-01"
[101] "2003-01-01" "2003-01-01" "1994-01-01" "1989-01-01" "1994-01-01"
[106] "1996-01-01" "1998-01-01" "1995-01-01" "1989-01-01" "1997-01-01"
[111] "1996-01-01" "1996-01-01" "1996-01-01" "1990-01-01" "2002-01-01"
[116] "2000-01-01" "1994-01-01" "1998-01-01"
```

Side-Note: Working with dates

We can use the lubridate package to ease the pain of doing math with dates.

```
library(lubridate)
```

```
Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

```
today()
```

```
[1] "2024-03-07"
```

```
today() - ymd("2002-11-19")
```

```
Time difference of 7779 days
```

```
time_length(today() - ymd("2002-11-19"), "years")
```

```
[1] 21.29774
```

So what is the age of everyone on our dataset.

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

> 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?

```
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
```

```
wPsubject <- filter(subject, infancy_vac == "wP")
head(wPsubject)
```

```
  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      age
1    1986-01-01    2016-09-12 2020_dataset 38.17933
2    1968-01-01    2019-01-28 2020_dataset 56.18070
3    1983-01-01    2016-10-10 2020_dataset 41.18001
4    1988-01-01    2016-08-29 2020_dataset 36.18070
5    1991-01-01    2016-08-29 2020_dataset 33.18001
6    1988-01-01    2016-10-10 2020_dataset 36.18070
```

```
mean(wPsubject$age)
```

```
[1] 36.57618
```

```
aPsubject <- filter(subject, infancy_vac == "aP")
mean(aPsubject$age)
```

```
[1] 26.27944
```

```
t.test(wPsubject$age, aPsubject$age)
```

```
    Welch Two Sample t-test

data:  wPsubject$age and aPsubject$age
t = 12.436, df = 65.411, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  8.643385 11.950080
sample estimates:
mean of x mean of y
 36.57618  26.27944
```

Average age of wP individuals is 36.58 yrs old and that of aP individuals is 26.28 yrs old. Yes they are significantly different with a p-value $<$ 2.2e-16.

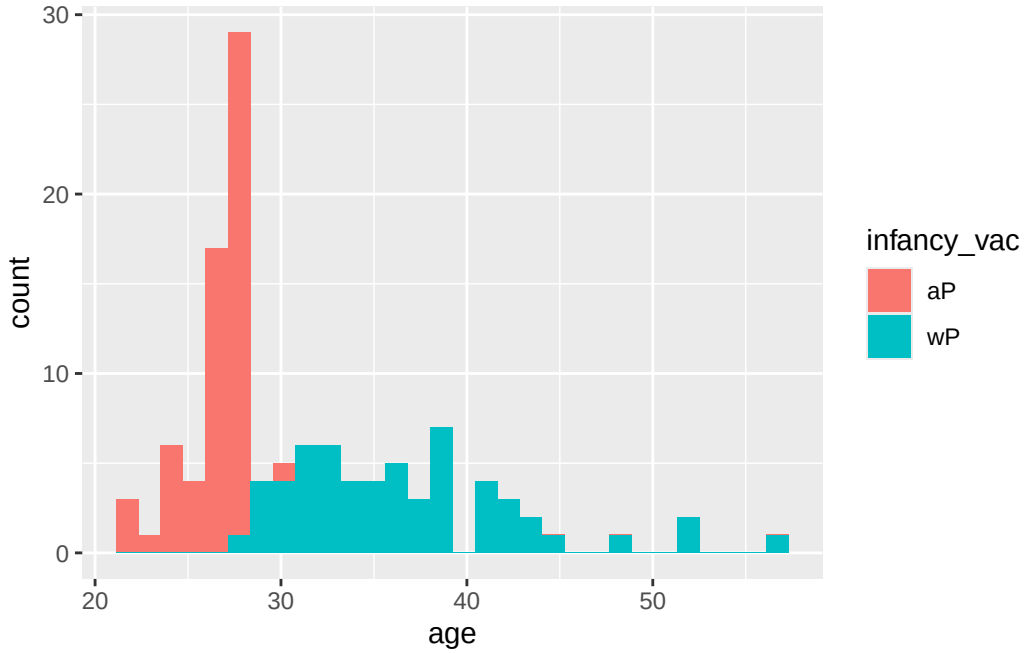
Q8. Determine the age of all individuals at time of boost?

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

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

```
ggplot(subject) +
  aes(age, fill = infancy_vac) +
  geom_histogram()
```

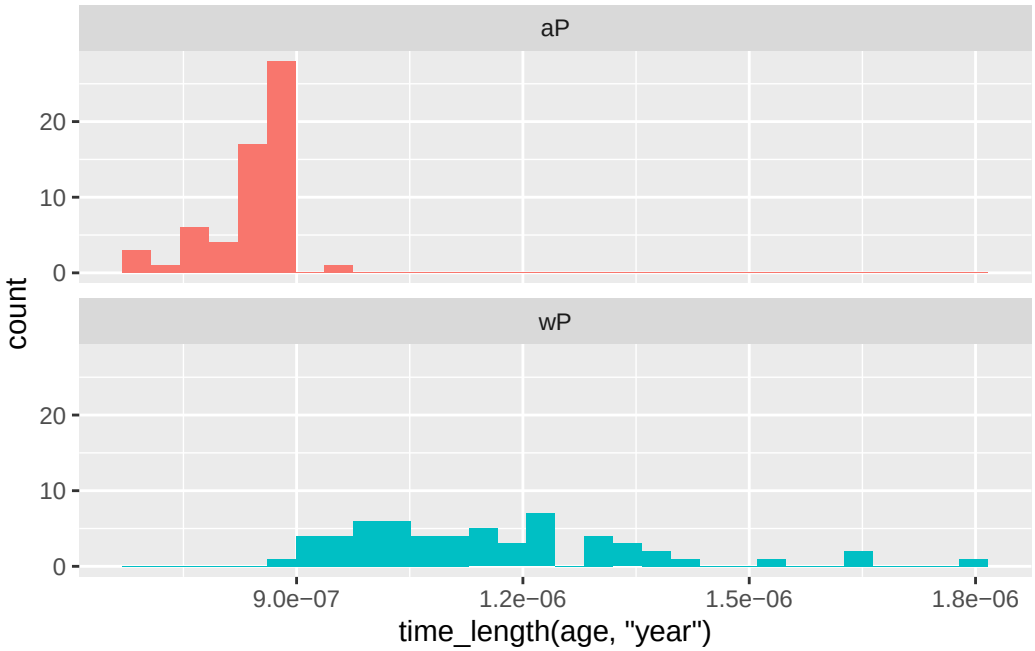
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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`.
```



Yes, those who received the wP vaccine seem to be significantly older than those who received the aP vaccine.

Get more data from CMI-PB

```
specimen <- read_json("http://cmi-pb.org/api/specimen", simplifyVector = T)
head(specimen)
```

```
  specimen_id subject_id actual_day_relative_to_boost
1           1          1                           -3
2           2          1                            1
3           3          1                            3
4           4          1                            7
5           5          1                           11
6           6          1                           32
  planned_day_relative_to_boost specimen_type visit
1                             0         Blood     1
2                             1         Blood     2
3                             3         Blood     3
4                             7         Blood     4
5                            14         Blood     5
```

```
6                              30         Blood     6
```

We need to **join** these two tables (subject and specimen) to make a single new "meta" table with all our metadata. We will use the `dplyr` join functions to do this:

> Q10. 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:

```
library(dplyr)

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

```
Joining with `by = join_by(subject_id)`
```

```
head(meta)
```

```
  subject_id infancy_vac biological_sex              ethnicity  race
1          1          wP         Female Not Hispanic or Latino White
2          1          wP         Female Not Hispanic or Latino White
3          1          wP         Female Not Hispanic or Latino White
4          1          wP         Female Not Hispanic or Latino White
5          1          wP         Female Not Hispanic or Latino White
6          1          wP         Female Not Hispanic or Latino White
  year_of_birth date_of_boost      dataset      age specimen_id
1    1986-01-01    2016-09-12 2020_dataset 38.17933           1
2    1986-01-01    2016-09-12 2020_dataset 38.17933           2
3    1986-01-01    2016-09-12 2020_dataset 38.17933           3
4    1986-01-01    2016-09-12 2020_dataset 38.17933           4
5    1986-01-01    2016-09-12 2020_dataset 38.17933           5
6    1986-01-01    2016-09-12 2020_dataset 38.17933           6
  actual_day_relative_to_boost planned_day_relative_to_boost specimen_type
1                           -3                             0         Blood
2                            1                             1         Blood
3                            3                             3         Blood
4                            7                             7         Blood
5                           11                            14         Blood
6                           32                            30         Blood
  visit
1     1
```

```
2     2
3     3
4     4
5     5
6     6
```

Now we can read some of the other data from CMI-PB

```
ab_titer <- read_json("http://cmi-pb.org/api/v4/plasma_ab_titer",
                      simplifyVector = T)

head(ab_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
```

One more `inner_join()` to add all our metadata in `meta` on to our `ab_data` table:

> Q11. 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.

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

```
Joining with `by = join_by(specimen_id)`
```

```
head(abdata)
```

```
  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 subject_id infancy_vac biological_sex
1 UG/ML                2.096133          1          wP         Female
2 IU/ML               29.170000          1          wP         Female
3 IU/ML                0.530000          1          wP         Female
4 IU/ML                6.205949          1          wP         Female
5 IU/ML                4.679535          1          wP         Female
6 IU/ML                2.816431          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 actual_day_relative_to_boost planned_day_relative_to_boost
1 38.17933                         -3                            0
2 38.17933                         -3                            0
3 38.17933                         -3                            0
4 38.17933                         -3                            0
5 38.17933                         -3                            0
6 38.17933                         -3                            0
  specimen_type visit
1         Blood     1
2         Blood     1
3         Blood     1
4         Blood     1
5         Blood     1
6         Blood     1
```

```
dim(abdata)
```

```
[1] 41775    21
```

Our first exploratory plot:

Q12. How many specimens (i.e. entries in abdata) do we have for each isotype?

```
table(abdata$antigen)
```

```
  ACT   BETV1      DT   FELD1     FHA  FIM2/3   LOLP1     LOS Measles     OVA
 1970    1970    3435    1970    3829    3435    1970    1970    1970    3435
  PD1     PRN      PT     PTM   Total      TT
 1970    3829    3829    1970     788    3435
```

Q13. What are the different $dataset values in abdata and what do you notice about the number of rows for the most "recent" dataset?

```
table(abdata$dataset)
```

```
2020_dataset 2021_dataset 2022_dataset
       31520         8085         2170
```

We have less rows for more recent years.

Examine IgG Ab titer levels

Now using our joined/merged/linked abdata dataset filter() for IgG isotype.

```
igg <- abdata %>% filter(isotype == "IgG")
head(igg)
```

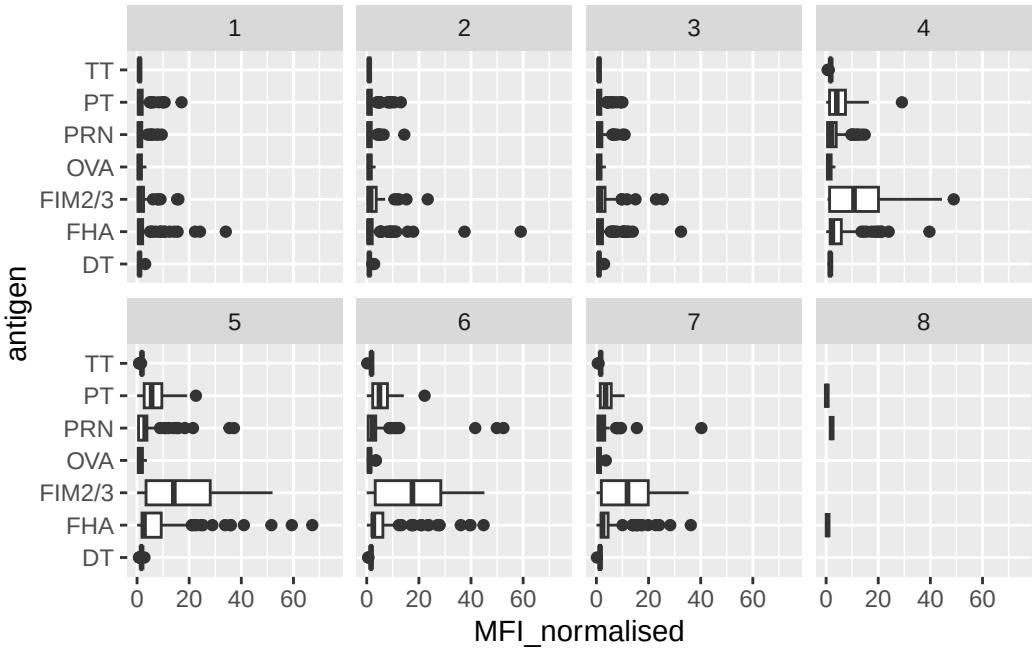
```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1     IgG                TRUE      PT   68.56614       3.736992
2           1     IgG                TRUE     PRN  332.12718       2.602350
3           1     IgG                TRUE     FHA 1887.12263      34.050956
4          19     IgG                TRUE      PT   20.11607       1.096366
5          19     IgG                TRUE     PRN  976.67419       7.652635
6          19     IgG                TRUE     FHA   60.76626       1.096457
   unit lower_limit_of_detection subject_id infancy_vac biological_sex
1 IU/ML                0.530000          1          wP         Female
2 IU/ML                6.205949          1          wP         Female
3 IU/ML                4.679535          1          wP         Female
```

```
4 IU/ML                 0.530000          3          wP         Female
5 IU/ML                 6.205949          3          wP         Female
6 IU/ML                 4.679535          3          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                 Unknown White    1983-01-01    2016-10-10 2020_dataset
5                 Unknown White    1983-01-01    2016-10-10 2020_dataset
6                 Unknown White    1983-01-01    2016-10-10 2020_dataset
       age actual_day_relative_to_boost planned_day_relative_to_boost
1 38.17933                          -3                             0
2 38.17933                          -3                             0
3 38.17933                          -3                             0
4 41.18001                          -3                             0
5 41.18001                          -3                             0
6 41.18001                          -3                             0
  specimen_type visit
1         Blood     1
2         Blood     1
3         Blood     1
4         Blood     1
5         Blood     1
6         Blood     1
```

Q14. Complete the following code to make a summary boxplot of Ab titer levels (MFI) for all antigens:

```
ggplot(igg) +
  aes(MFI_normalised, antigen) +
  geom_boxplot() +
    xlim(0,75) +
  facet_wrap(vars(visit), nrow=2)
```

```
Warning: Removed 5 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```

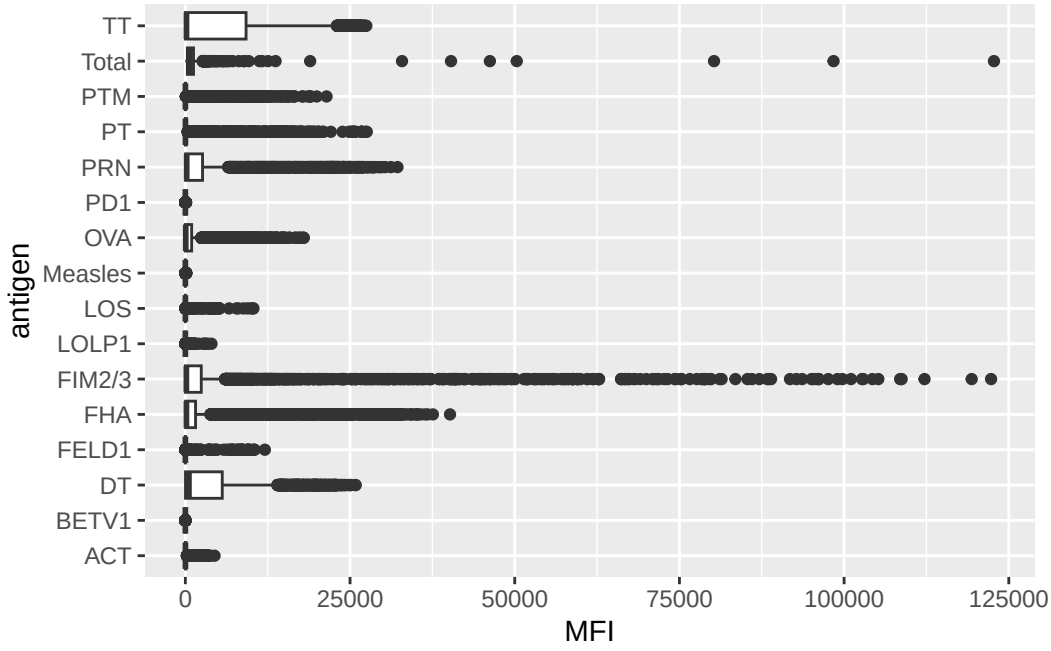


Q15. What antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others?

PT, PRN, FIM2/3, and FHA. These were included in the vaccine.

```
ggplot(abdata) +
  aes(MFI, antigen) +
  geom_boxplot()
```

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_boxplot()`).
```

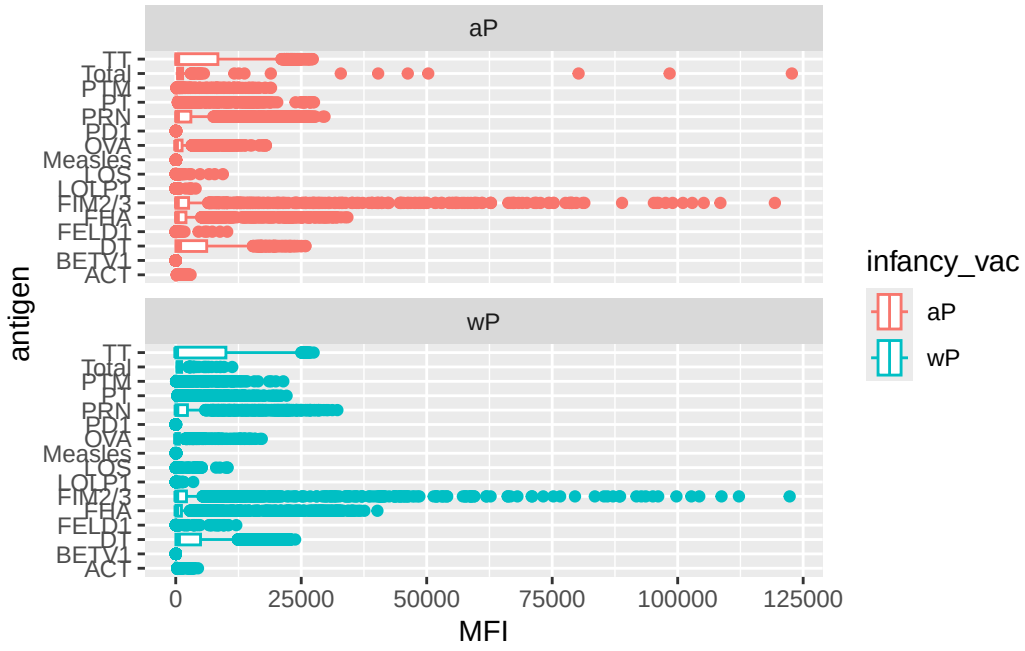


Why are certain antigens and not others very variable in their detected levels here? The vaccine only contains certain antigens.

Can you facet or even just color by infancy_vac? Is there some difference?

```
ggplot(abdata) +
  aes(MFI, antigen, col=infancy_vac) +
  geom_boxplot() +
  facet_wrap(vars(infancy_vac), nrow=2)
```

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_boxplot()`).
```



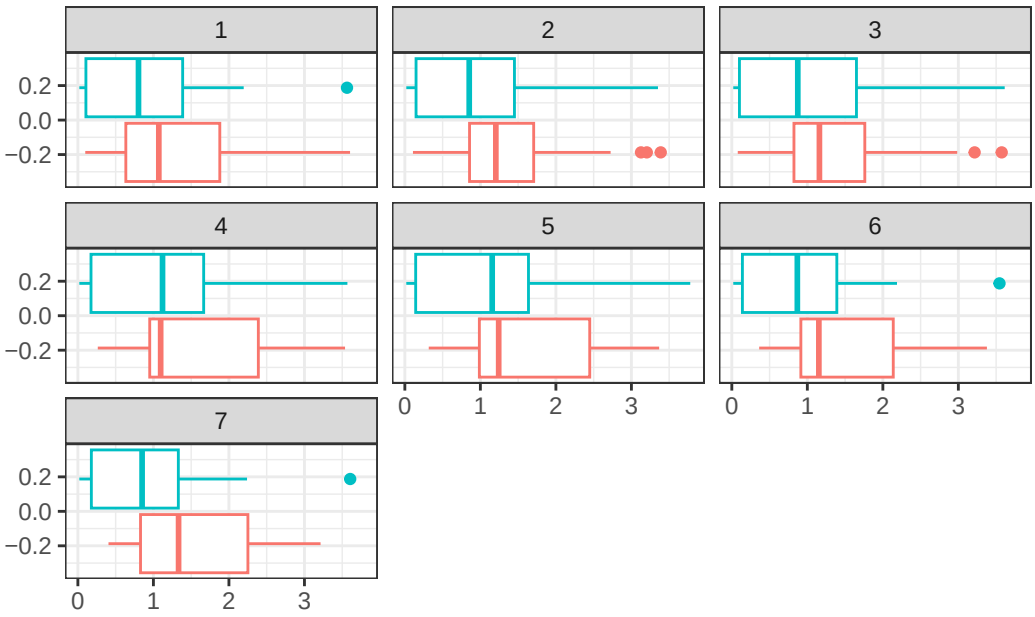
There are potentially some differences here but in general it is hard to tell with this whole dataset overview...

```
table(abdata$dataset)
```

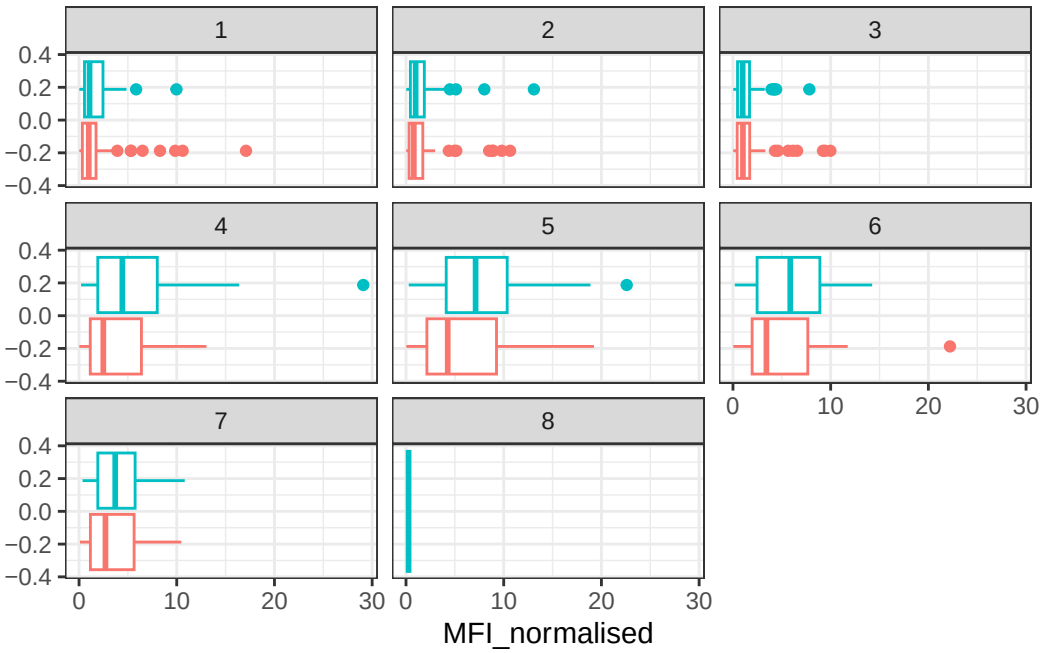
```
2020_dataset 2021_dataset 2022_dataset
       31520         8085         2170
```

Q16. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("OVA", that is not in our vaccines) and a clear antigen of interest ("PT", Pertussis Toxin, one of the key virulence factors produced by the bacterium B. pertussis).

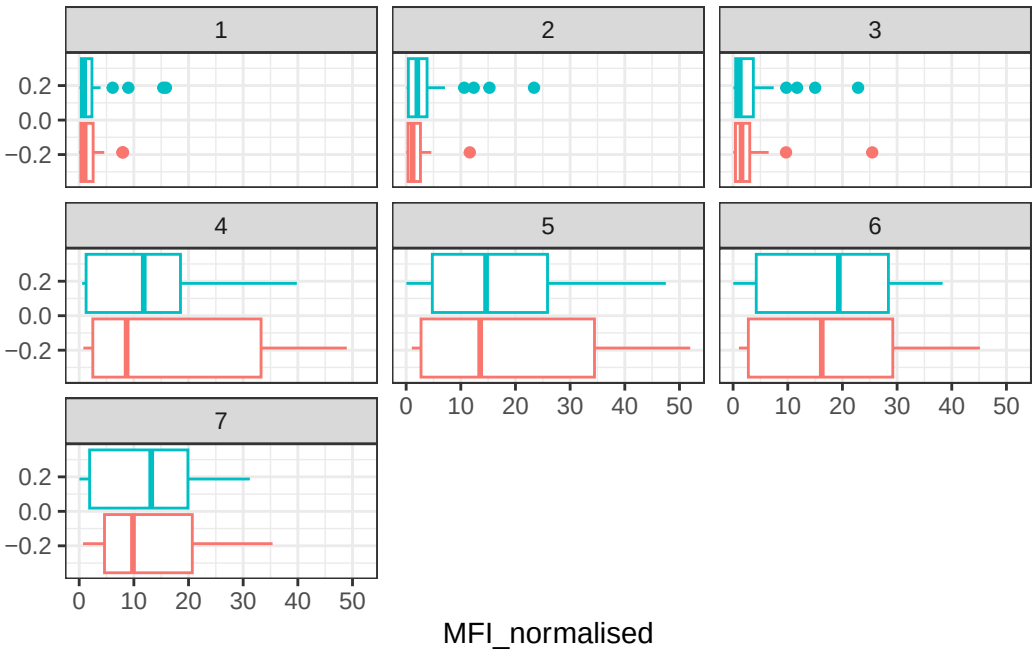
```
filter(igg, antigen=="OVA") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = F) +
  facet_wrap(vars(visit)) +
  theme_bw()
```




```
filter(igg, antigen=="PT") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = F) +
  facet_wrap(vars(visit)) +
  theme_bw()
```




```
filter(igg, antigen=="FIM2/3") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = F) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



Q17. What do you notice about these two antigens time courses and the PT data in particular?

PT increases for a while then has a sharp drop off while FIM 2/3 steadily increases over time and only slightly decreases.

Let's focus in on just the 2021_dataset.

```
abdata.21 <- filter(abdata, dataset == "2021_dataset")
table(abdata.21$dataset)
```

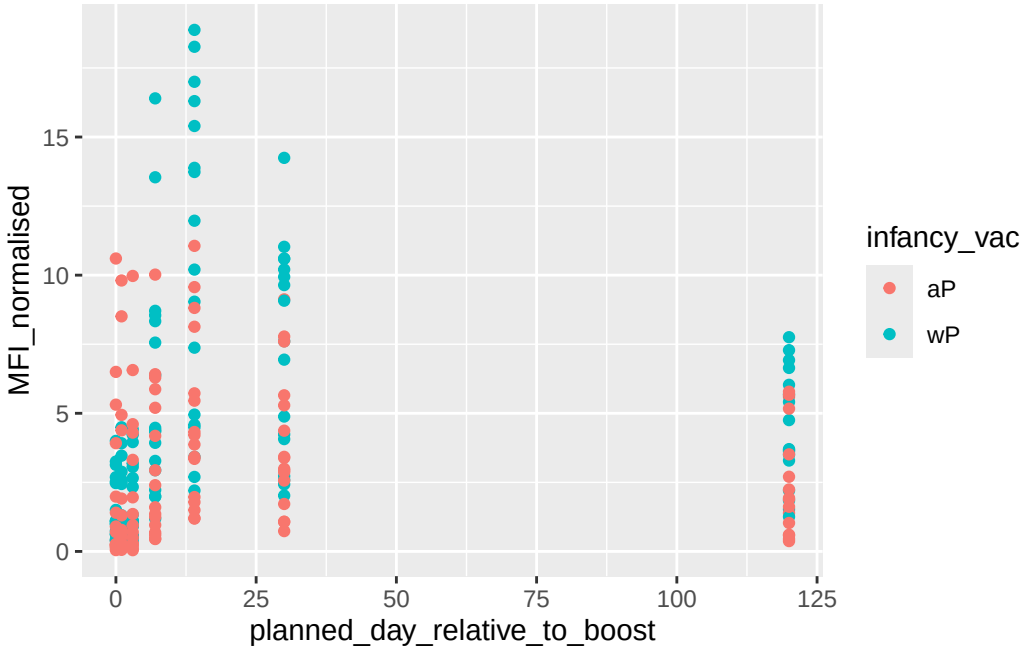
```
2021_dataset
        8085
```

Focus on PT antigen for IgG levels

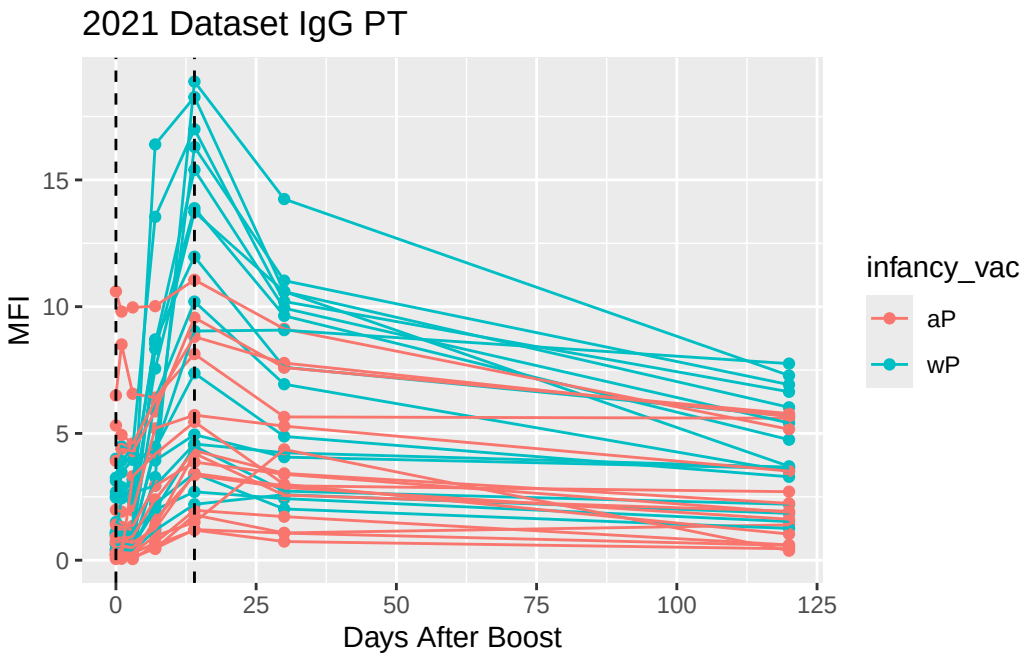
```
pt.21 <- filter(abdata.21, isotype == "IgG", antigen == "PT")
```

Plot of

```
ggplot(pt.21) +
  aes(x=planned_day_relative_to_boost,
        y=MFI_normalised,
        col=infancy_vac,) +
    geom_point()
```




```
ggplot(pt.21) +
  aes(x=planned_day_relative_to_boost,
        y=MFI_normalised,
        col=infancy_vac,
        group=subject_id) +
    geom_point() +
    geom_line() +
    geom_vline(xintercept=0, linetype="dashed") +
    geom_vline(xintercept=14, linetype="dashed") +
    labs(title="2021 Dataset IgG PT", x = "Days After Boost", y = "MFI")
```



Q18. Do you see any clear difference in aP vs. wP responses?

wP responses show a greater peak at 14-days post-vaccination.