class 15

Mia Fava

##Background

Pertussis, aka whooping coughm is a highly infectious lunch disease caused by the bacteria *B. pertussius*

THe CDC tracks pertussius cases number per year CDC data

##Section 2:

We will use the **datapasta** R package to "scrape" this data to R

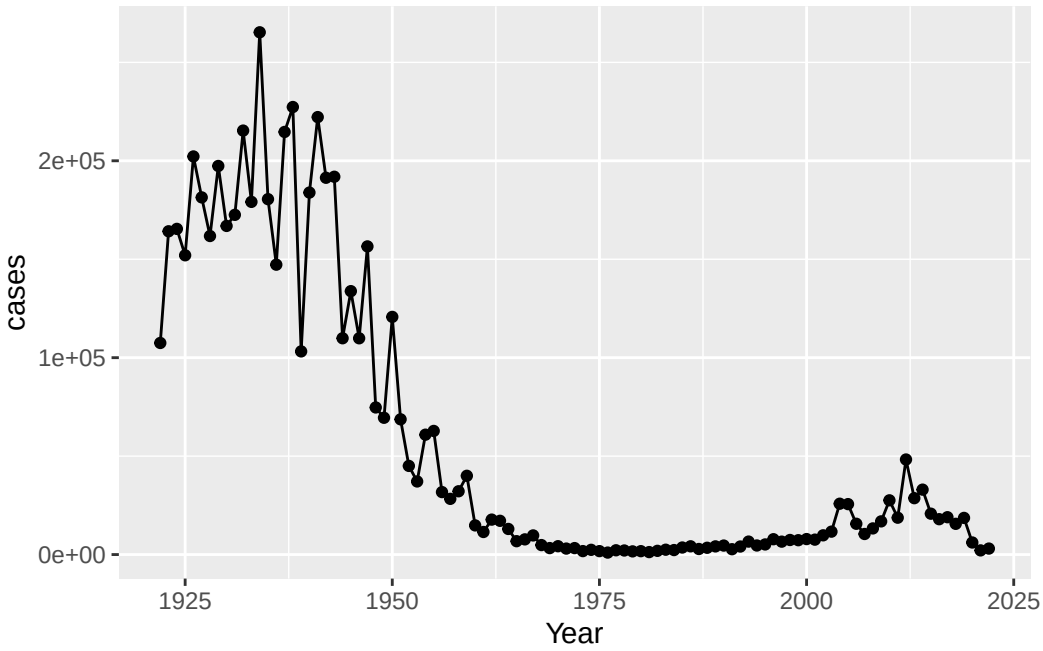
```
cdc <- data.frame(
                                    Year = c(1922L,1923L,1924L,1925L,
                                             1926L,1927L,1928L,1929L,1930L,1931L,
                                             1932L,1933L,1934L,1935L,1936L,
                                             1937L,1938L,1939L,1940L,1941L,1942L,
                                             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,2020L,2021L,2022L),
        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,
                                        6124,2116,3044)
      )
```

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)
```

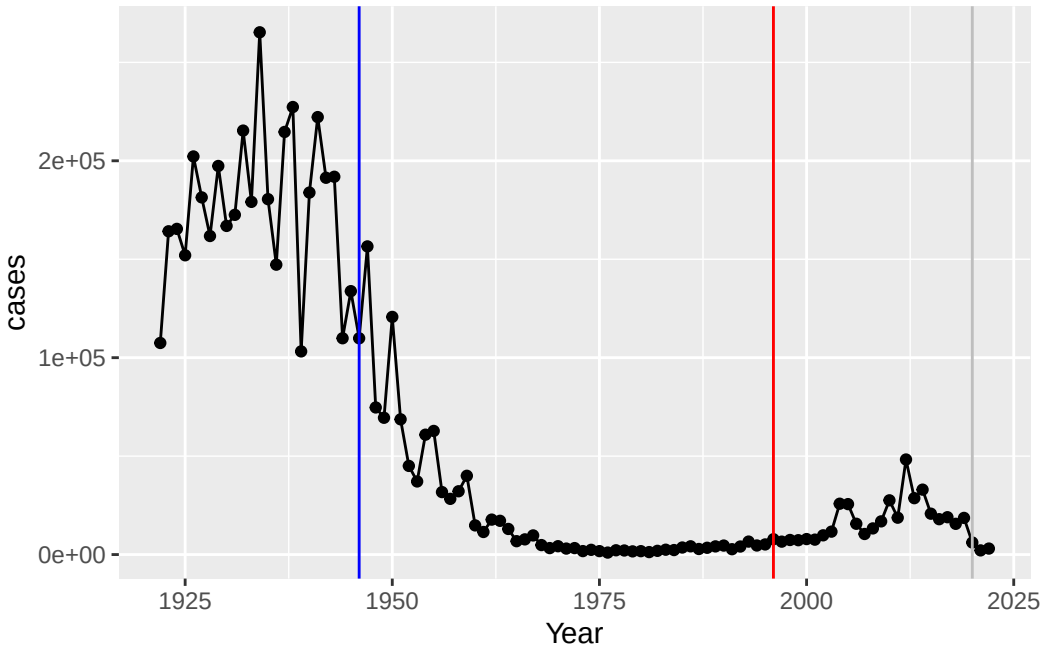
```
ggplot(cdc) +
  aes(Year, cases) +
  geom_point() +
  geom_line()
```



Add some landmark developments as annotation to our plot. We include the first whole-cell (wP) vaccine roll-out in 1940.

> 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?

```
ggplot(cdc) +
  aes(Year, cases) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept=1946, col="blue") +
  geom_vline(xintercept=1996, col="red") +
  geom_vline(xintercept=2020, col="grey")
```



wP=blue line ; aP= red line ; COVID = grey line

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

We went from ~200,000 cases to ~1,000 cases in 1976, but when switched to the aP vaccine there was a significant increase ~10 year later. This increase could be due to individuals being more hesitant to take the vaccine (anti vaxers) or evolution of the bacteria against the antibiotic – more resistance to the vaccine and new variants showing up.

Key question : Why does the aP vaccine induce immunity wane faster than that of the wP vaccine?

##Section 3

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

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

```
aP wP
87 85
```

There are 87 aP vaccinated subjects and 85 wP vaccinated subjects.

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

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

```
Female   Male
   112     60
```

There are 112 females and 60 males in the dataset.

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                                     32   12
  Black or African American                  2    3
  More Than One Race                        15    4
  Native Hawaiian or Other Pacific Islander  1    1
  Unknown or Not Reported                   14    7
  White                                     48   32
```

The breakdown of race and biological sex is seen above. Majority of the data set seems to be Asian or White, predominantly females. This dataset is not a good representation of the United States.

> 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(lubridate)
```

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

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

    date, intersect, setdiff, union
```

> the average age of wP individuals

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

```
  [1] 14204 20779 15300 13474 12378 13474 16030 14569 10552 15665 14204 15665
 [13] 10186 11647 13108 13839 16396 10186 11282 16030 15300 14569 12378 12013
 [25] 13474 15300 10186 15665 10186 13474 13108 10186 12743 15300 12378 10186
 [37]  9821 10186 14569 11282 14569 10186  9821  9821 10186  9821 10552  9821
 [49] 10186 10186 10186  9821  9821 10186 10186 10186 10552 10186 10186 10186
 [61] 13839 11647 10917 11647 12743 17857 19318 19318 12743  9821  9821 12378
 [73] 10917 10917  9821  9821 13474 11647 13839 12013 11647  9821  9456 10186
 [85]  9091  9821  9091  9091 10186  9456  9821  9091 10552  9456  9821  9091
 [97] 14204 11647  9456  8725  7995  7995 11282 13108 11282 10552  9821 10917
[109] 13108 10186 10552 10552 10552 12743  8360  9091 11282  9821  9821 10917
```

```
 [121]  9091  9456 10552  9091 11647 11647 10552 11282 12378 10552  9821 10917
 [133] 10186 12743 10917 10917  9821  9091 11647  8725 10552 12378  7995  9456
 [145]  8360 12013  9091 13474 12378 12378 12013 10917  9821 10186 10186  8725
 [157] 10186  9091 11282 10552 11647  9456 11647 12378 11647  8725 10186 12378
 [169]  7995 12013  7995 14204
```

```
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 <- subject %>% filter(infancy_vac == "aP")
round( summary( time_length( ap$age, "years" ) ) )
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     22      26      27      27      28      34
```

```
wp <- subject %>% filter(infancy_vac == "wP")
round( summary( time_length( wp$age, "years" ) ) )
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     22      32      34      36      39      57
```

The data is signficantly different based off of the tests above, but to double check – let's perform a p-test. $p<0.05$ then it is significant

```
x <- t.test(time_length( wp$age, "years" ),
       time_length( ap$age, "years" ))

x$p.value
```

```
[1] 2.372101e-23
```

The value is 2.37e-23, therefore the data is significantly different.

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

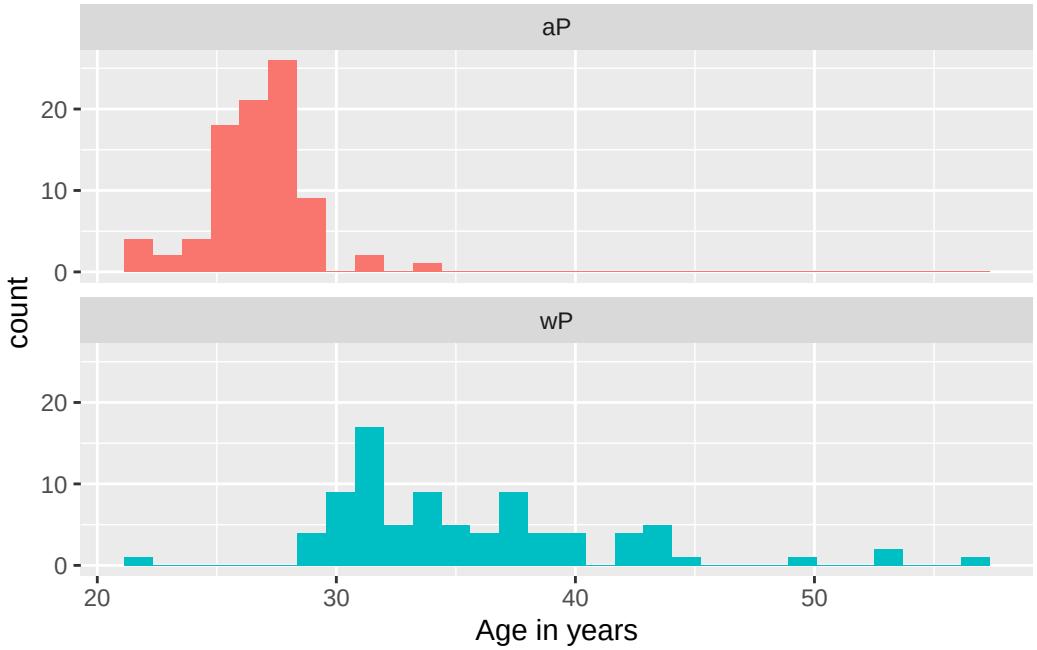
```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

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

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) +
  xlab("Age in years")
```

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



Yes, they are significantly different which is evident in their graphs. the aP data is seen to be focused more around below the ages of 30, whereas the wP dataset is spread more across the age range of 30-45.

##Joining multiple tables

```
specimen <- read_json("https://www.cmi-pb.org/api/v5/specimen", simplifyVector = TRUE)
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
```

Now we can merge the two tables together `subject` and `specimen` to make one new `meta` table with the combined data.

> 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:

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

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

```
head(meta)
```

```
  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 infancy_vac biological_sex
1                             0         Blood     1          wP         Female
2                             1         Blood     2          wP         Female
3                             3         Blood     3          wP         Female
4                             7         Blood     4          wP         Female
5                            14         Blood     5          wP         Female
6                            30         Blood     6          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 14204 days
2 14204 days
3 14204 days
4 14204 days
5 14204 days
6 14204 days
```

Now read the "experiment data" table from CMI-PB

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.

```
library(jsonlite)
abdata <- read_json("https://www.cmi-pb.org/api/v5/plasma_ab_titer", simplifyVector = TRUE)
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
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 merge to do of `meta` with `abdata` to associate all metadata about the individual and their race, biological sex, and infancy vaccination status together with Antibody levels…

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

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

```
head(ab)
```

```
  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 actual_day_relative_to_boost
1 UG/ML                 2.096133          1                           -3
2 IU/ML                29.170000          1                           -3
3 IU/ML                 0.530000          1                           -3
4 IU/ML                 6.205949          1                           -3
5 IU/ML                 4.679535          1                           -3
6 IU/ML                 2.816431          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 14204 days
2 14204 days
3 14204 days
4 14204 days
5 14204 days
6 14204 days
```

How many Ab measurements do we have?

```
nrow(ab)
```

```
[1] 52576
```

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

```
table(ab$isotype)
```

```
  IgE   IgG  IgG1  IgG2  IgG3  IgG4
 6698  5389 10117 10124 10124 10124
```

Q12. What are the different $dataset values in abdata and what do you notice about the number of rows for the most "recent" dataset?

```
dataset_table <- table(ab$dataset)
print(dataset_table)
```

```
2020_dataset 2021_dataset 2022_dataset 2023_dataset
       31520         8085         7301         5670
```

The dataset values in ab seem to be at a peak in 2021, but then decrease in the following years. In 2020, the values were the lowest! The most recent set could have lower numbers due to how it is still being built or haven't had every value collected.

How many antigens

```
table(ab$antigen)
```

```
    ACT   BETV1      DT   FELD1     FHA  FIM2/3   LOLP1     LOS Measles     OVA
   1970    1970    4978    1970    5372    4978    1970    1970    1970    4978
    PD1     PRN      PT     PTM   Total      TT
   1970    5372    5372    1970     788    4978
```

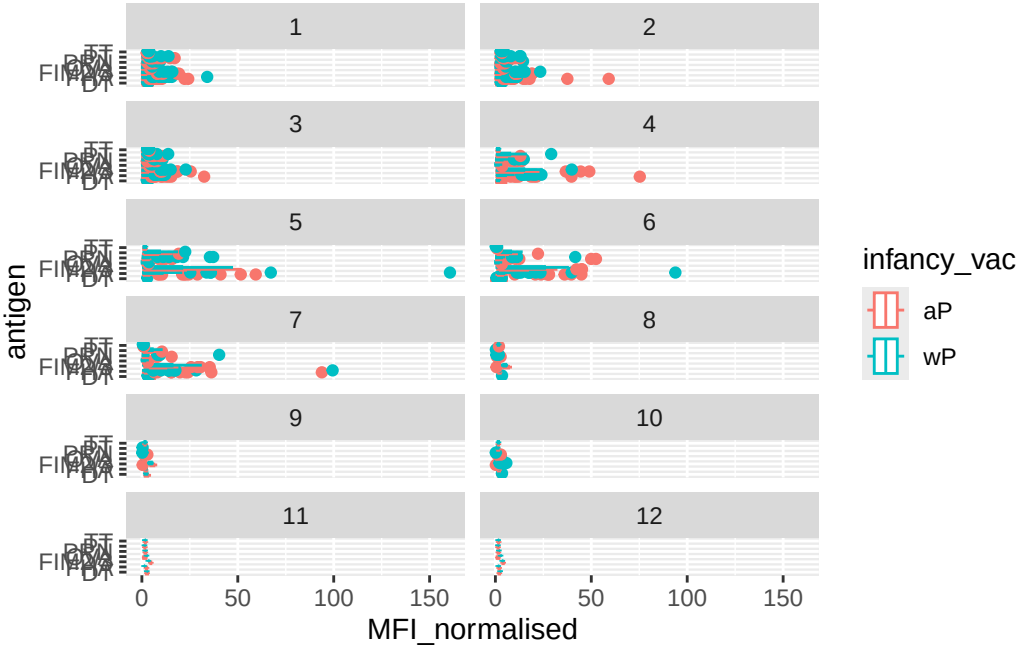
##Section 4 Make a plot of MFI (measure of how much is detected)

```
igg <- filter(ab, 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 actual_day_relative_to_boost
1 IU/ML                 0.530000          1                           -3
2 IU/ML                 6.205949          1                           -3
3 IU/ML                 4.679535          1                           -3
4 IU/ML                 0.530000          3                           -3
5 IU/ML                 6.205949          3                           -3
6 IU/ML                 4.679535          3                           -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                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
1 14204 days
2 14204 days
3 14204 days
4 15300 days
5 15300 days
6 15300 days
```

```
ggplot(igg) +
  aes(MFI_normalised, antigen, col=infancy_vac)+
  geom_boxplot()+
  facet_wrap(~visit, ncol=2)
```




```
table(igg$visit)
```

```
  1   2   3   4   5   6   7   8   9  10  11  12
902 902 930 559 559 540 525 150 147 133  21  21
```

Only looking at first 8 visits

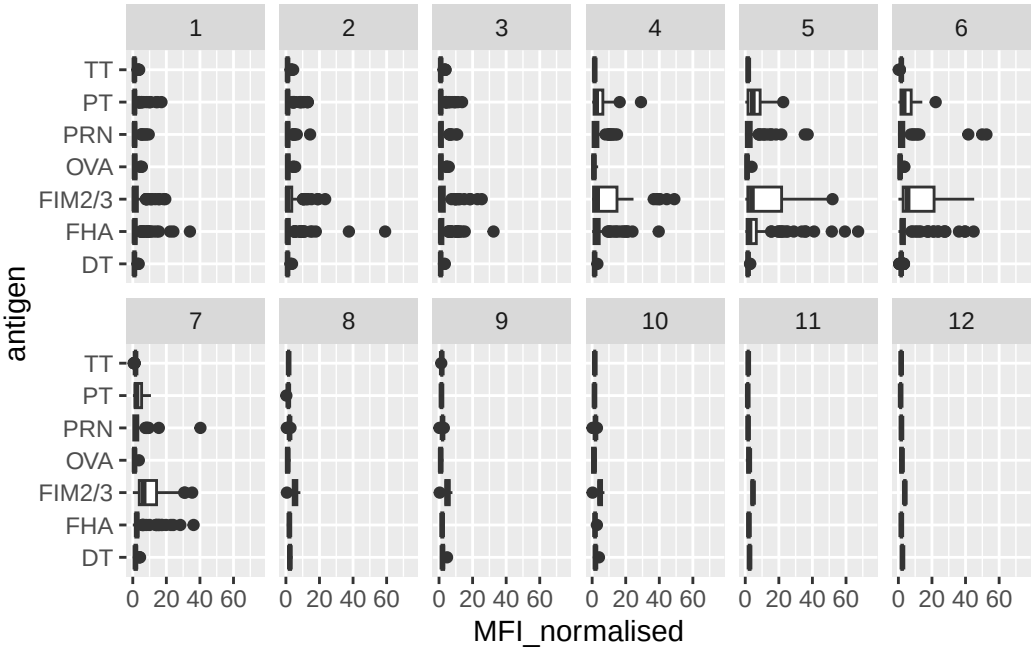
```
igg_7 <- filter(igg, visit %in% 1:7)
table(igg_7$visit)
```

```
  1   2   3   4   5   6   7
902 902 930 559 559 540 525
```

Q13. 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()`).
```

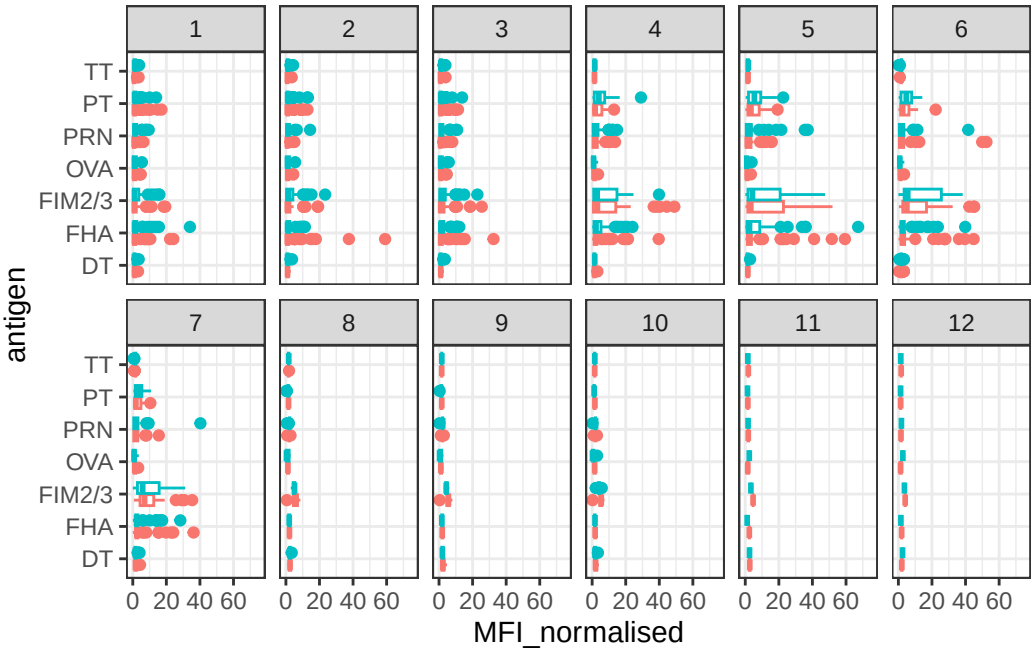


Q14. What antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others?

The antigens that show differences in the level of IgG antibody titers include the FIM2/3, PRN, PT, and the FHA. This makes sense due to how these antigens are key targets for immunity in the vaccines for pertussis. There is an evident difference in level of IgG antibody over time from the plots, as seen when comparing each visits.

```
ggplot(igg) +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  xlim(0,75) +
  theme_bw()
```

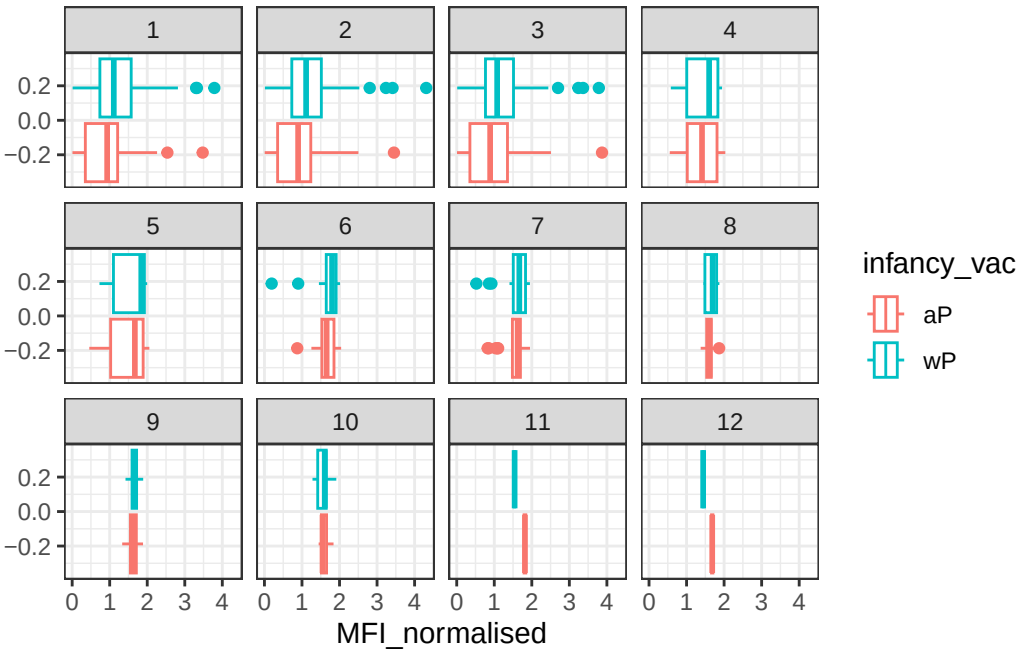
```
Warning: Removed 5 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



Q15. 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).

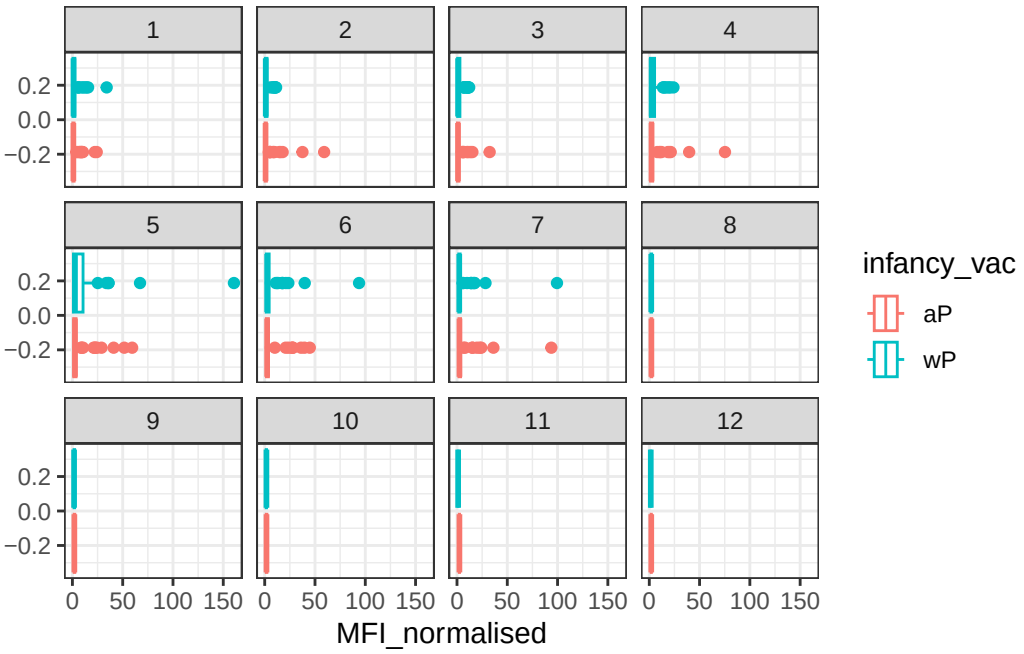
TT antigen levels per visit

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



PRN anitgen levels per visit

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



Q16. What do you notice about these two antigens time courses and the PT data in particular?

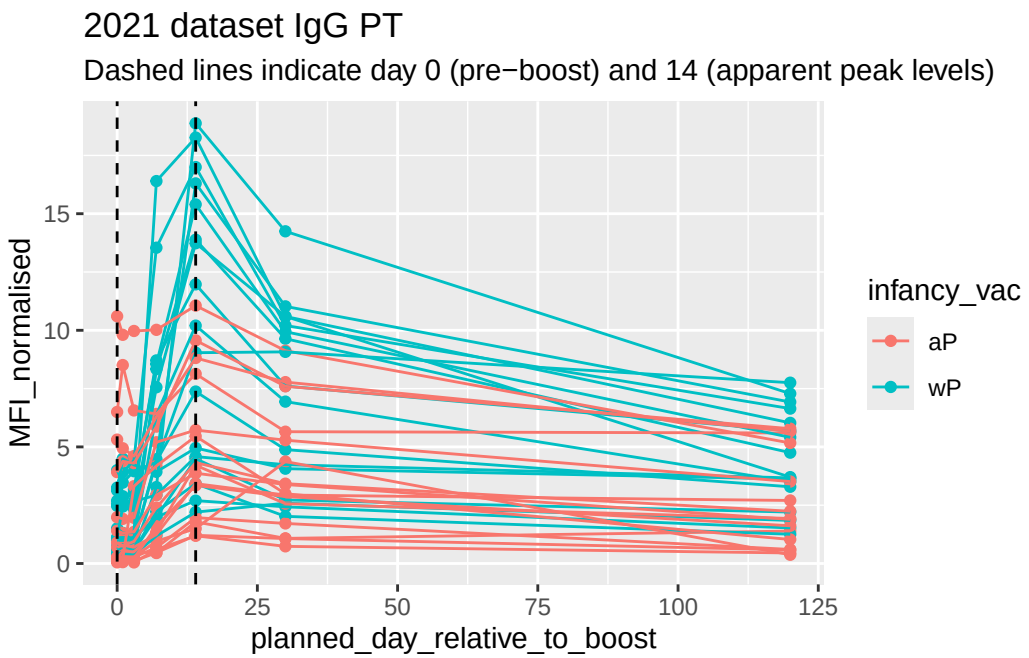
The levels in the two antigens I picked, TT and FHA, it is obvious that the levels in FHA exceed those of TT. The levels of FHA are seen to really peak around visit 6 and then decline after that in the following visits. The trend is similar between both aP and wP individuals.

Q17. Do you see any clear difference in aP vs. wP responses?

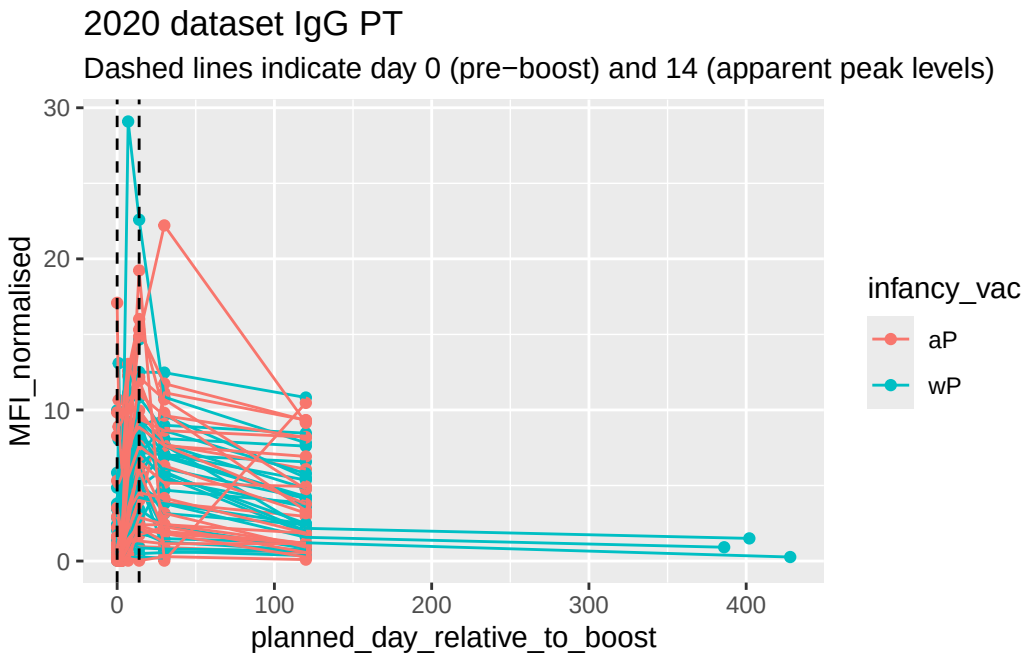
In regards to aP and wP individuals, there is not a vast difference between the two. They seem to be pretty consistent with each other across both antigens and have similar distributions.

```
abdata.21 <- ab %>% filter(dataset == "2021_dataset")

abdata.21 %>%
  filter(isotype == "IgG",  antigen == "PT") %>%
  ggplot() +
    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",
       subtitle = "Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)")
```




```
abdata.21 <- ab %>% filter(dataset == "2020_dataset")

abdata.21 %>%
  filter(isotype == "IgG",  antigen == "PT") %>%
  ggplot() +
    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="2020 dataset IgG PT",
       subtitle = "Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)")
```


2020 dataset IgG PT
Dashed lines indicate day 0 (pre−boost) and 14 (apparent peak levels)
MFI_normalised
30
20
10
0
0
100
200
300
400
planned_day_relative_to_boost
infancy_vac
aP
wP