Class 18: Pertussis Mini-project

Hyejeong Choi (PID: A16837133)

Table of contents

Background

Pertussis (a.k.a. whooping cough) is a common lung infection caused by the bacteria *B. Pertussis*.

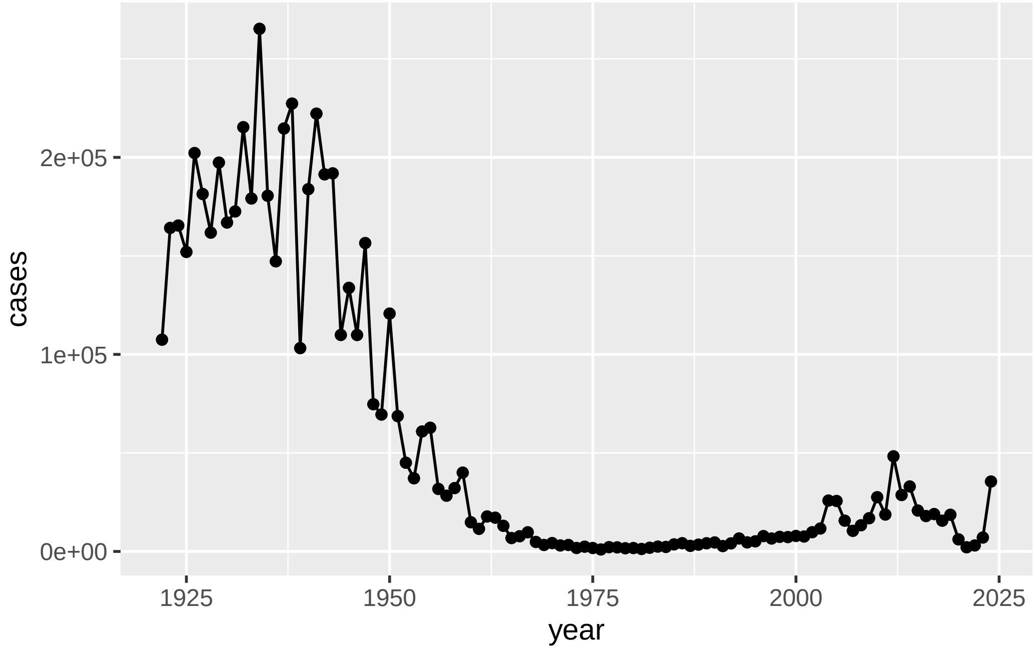
The CDC tracks cases of Pertussis in the US: https://tinyurl.com/pertussiscdc

Examining cases of Pertussis by year

We can use the **datapasta** package to scrape case numbers from the 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.

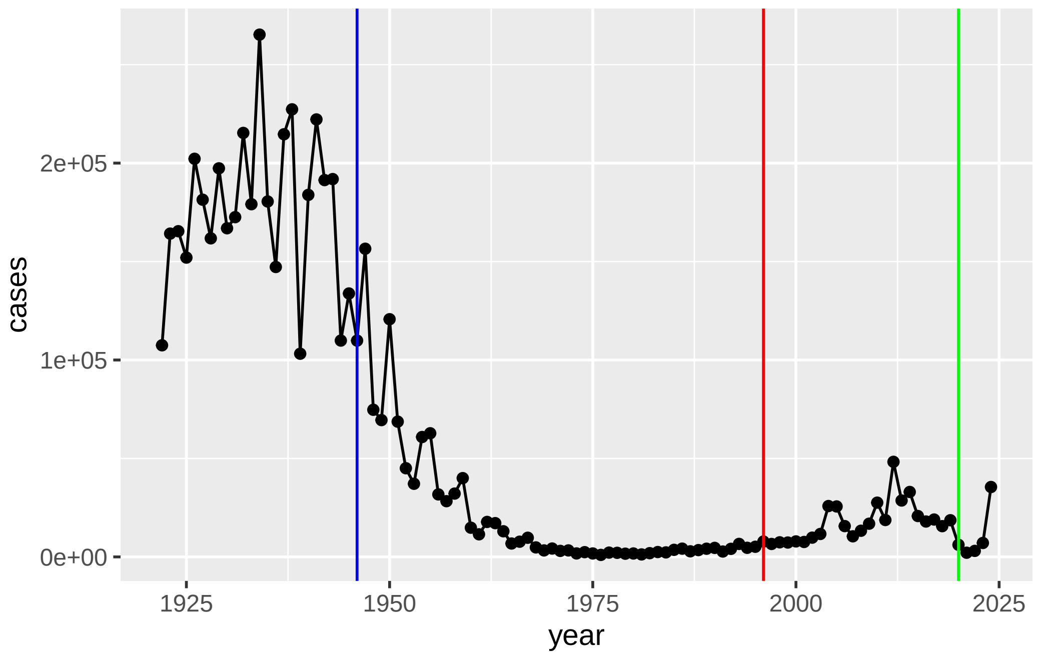
```
library(ggplot2)

cases <- ggplot(cdc) +
  aes(year, cases) +
  geom_line() +
  geom_point()

cases
```



> Q2. Add some key time points in our history of interaction with Pertussis. These include wP roll-out (the first vaccine) in 1946 and the switch to aP in 1996.

We can use `geom_vline()` for this.

```
cases +
  geom_vline(xintercept = 1946, col="blue") +
  geom_vline(xintercept = 1996, col="red") +
  geom_vline(xintercept = 2020, col="green")
```



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

The wP vaccine was very effective at reducing the number of Pertussis cases. After the aP vaccine, the number of cases increased. This may be due to evolution of the bacteria or a reduced number of people getting the vaccinations. The immune protection from the aP vaccine may also fade faster than the wP vaccine.

Mounting evidence suggests that the newer **aP** is less effective over the long term than the older **wP** vaccine that is replaced. In other words, vaccine protection wanes more rapidly with aP than with wP.

Enter the CMI-PB Project

CMI-PB (Computational Models of Immunity - Pertussis boost) major goal is to investigate how the immune system responds differently with aP vs wP vaccinated individuals and be able to predict this at an early stage.

CMI-PB makes all their collected data freely available and they store it in a database composed of different tables. Here we will access a few of these.

We can use the **jsonlite** package to read this data.

```
library(jsonlite)

subject <- read_json("https://www.cmi-pb.org/api/v5_1/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
```

> Q. How many subjects (i.e. enrolled people) are there in this dataset?

```
nrow(subject)
```

```
[1] 172
```

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

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

```

aP wP 
87 85 
```

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

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

```
Female   Male 
   112     60 
```

> 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
```

> Q. Is this representative of the US population?

No, this is not representative. This is representative of the UCSD students population because the majority of the data was taken from the students.

Working with dates

```
library(lubridate)
```

> 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?

```
# subtract subject date of birth from today to find 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
```

```
# find the age in years of the subjects in the aP vaccine group

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 
```

The average age of aP individuals is 27 years.

```
# find the age in years of the subjects in the wP vaccine group

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 average age of the wP individuals is 36 years.

Yes, the ages between the two groups are significantly different. The wP individuals are on average about 10 years older than the aP individuals.

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

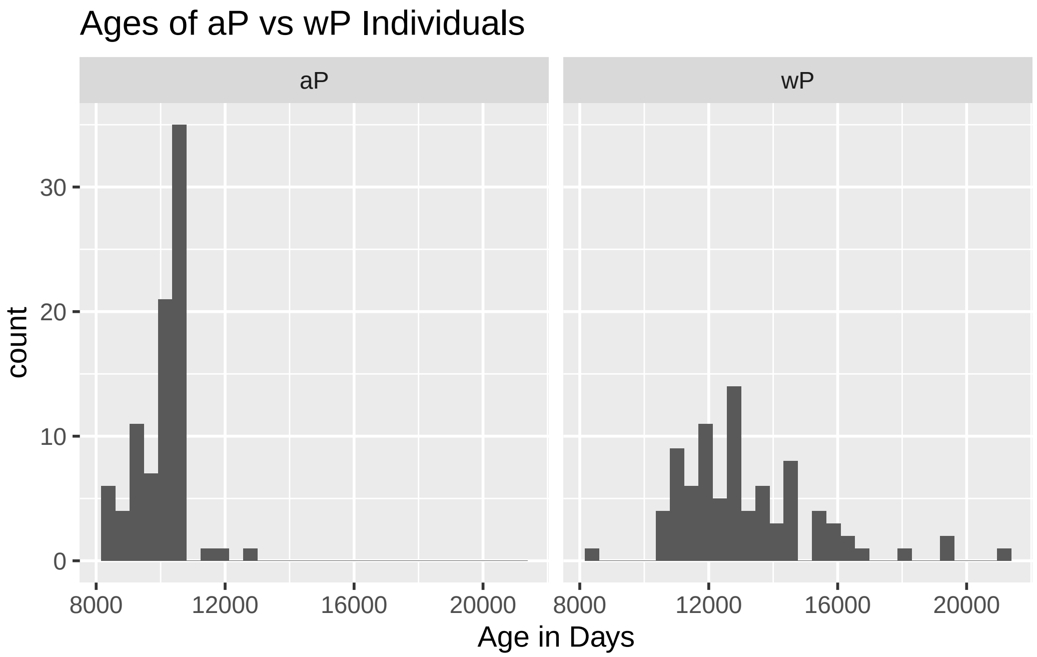
```r
# subtract day of birth from day of boost to get age at boost

boost_age_days <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)

# convert the age into years

boost_age_years <- time_length(boost_age_days, 'years')
head(boost_age_years)
```

```
[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?

```r
ggplot(subject) +
  aes(age) +
  geom_histogram() +
  facet_wrap(~infancy_vac) +
  labs(title='Ages of aP vs wP Individuals', x='Age in Days')
```

```
Don't know how to automatically pick scale for object of type <difftime>.
Defaulting to continuous.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Yes, there is a significant difference in the ages of the aP and wP groups.

Joining Multiple Tables

```
specimen <- read_json("https://www.cmi-pb.org/api/v5_1/specimen",
                      simplifyVector = TRUE)

ab_data <- read_json("https://www.cmi-pb.org/api/v5_1/plasma_ab_titer",
                     simplifyVector = TRUE)
```

Look at these data:

```
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 want to "join" these tables to get all our information together. For this, we will use the **dplyr** package and the `inner_join()` function.

> 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:

```
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 14394 days           1
2    1986-01-01    2016-09-12 2020_dataset 14394 days           2
3    1986-01-01    2016-09-12 2020_dataset 14394 days           3
4    1986-01-01    2016-09-12 2020_dataset 14394 days           4
5    1986-01-01    2016-09-12 2020_dataset 14394 days           5
6    1986-01-01    2016-09-12 2020_dataset 14394 days           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
```

> 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.

One more "join" to get ab_data and meta all together

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

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

```r
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 14394 days                           -3                             0
2 14394 days                           -3                             0
3 14394 days                           -3                             0
4 14394 days                           -3                             0
5 14394 days                           -3                             0
6 14394 days                           -3                             0
  specimen_type visit
1         Blood     1
2         Blood     1
3         Blood     1
4         Blood     1
5         Blood     1
6         Blood     1
```

```r
dim(abdata)
```

```
[1] 61956    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  7265 11993 12000 12000 12000 
```

Q. How many different antigens are measured in the dataset?

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

```

    ACT   BETV1      DT   FELD1     FHA  FIM2/3   LOLP1     LOS Measles     OVA 
   1970    1970    6318    1970    6712    6318    1970    1970    1970    6318 
    PD1     PRN      PT     PTM   Total      TT 
   1970    6712    6712    1970     788    6318 
```

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

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

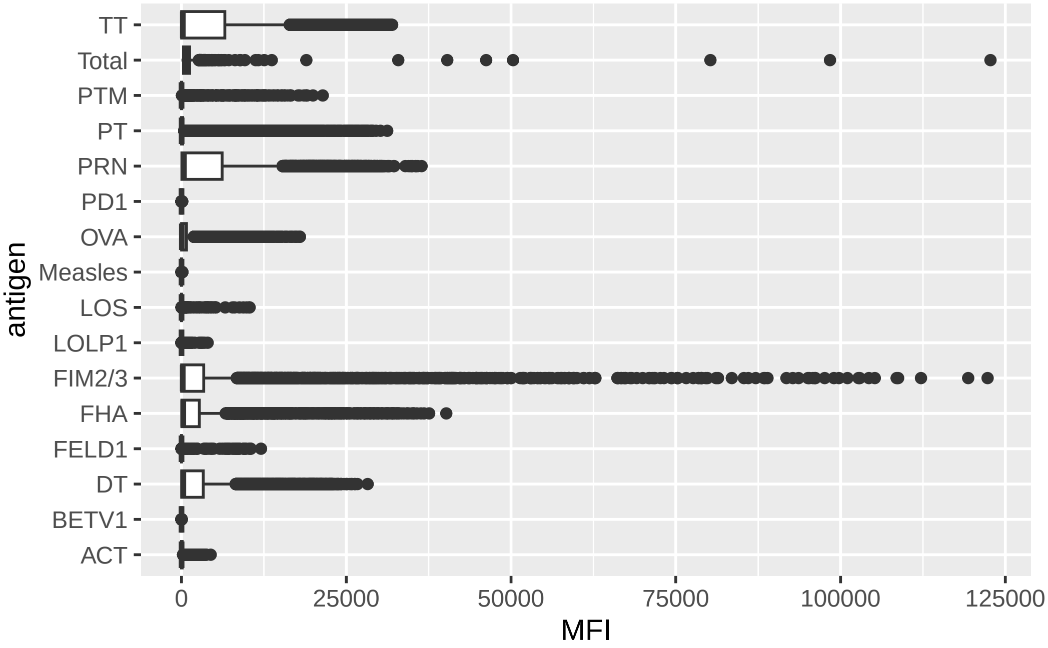
```

2020_dataset 2021_dataset 2022_dataset 2023_dataset 
       31520         8085         7301        15050 
```

The different values for $dataset are from the years 2020-2023. The most recent dataset has half the number of rows from the 2020 dataset, but about double the rows than the 2021 and 2022 datasets.

Q. Make a boxplot of antigen levels across the whole dataset (MFI vs antigen)

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

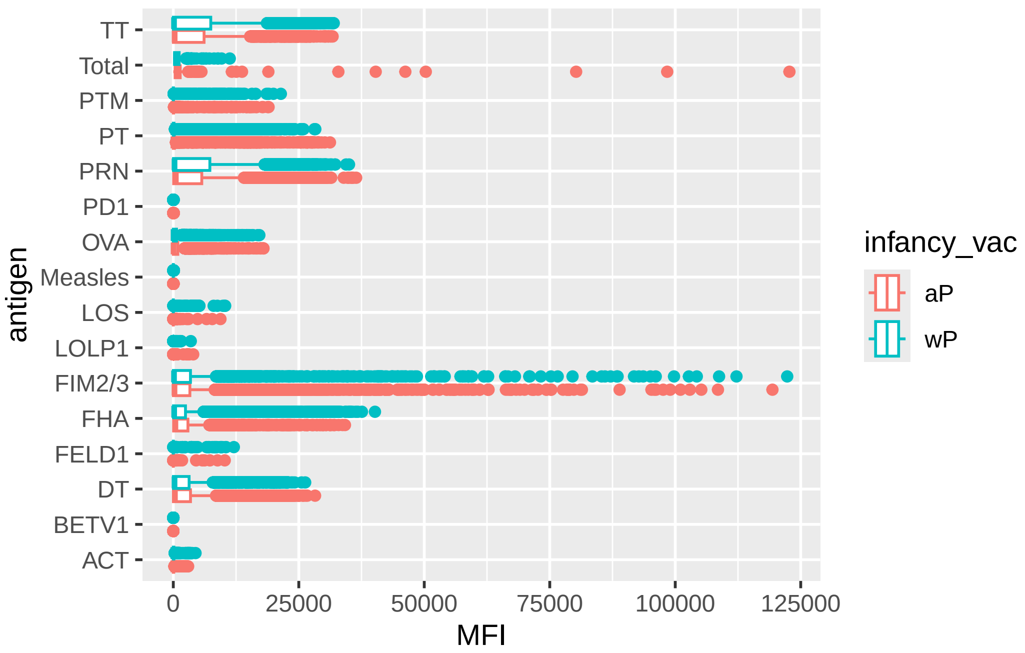
```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_boxplot()`).
```



Q. Are there obvious differences between aP and wP values?

```
ggplot(abdata) +
  aes(MFI, antigen, col=infancy_vac) +
  geom_boxplot()
```

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



Focus on IgG levels

IgG is the most abundant antibody in blood. With four sub-classes (IgG1 to IgG4) crucial for long-term immunity and responding to bacterial and viral infections.

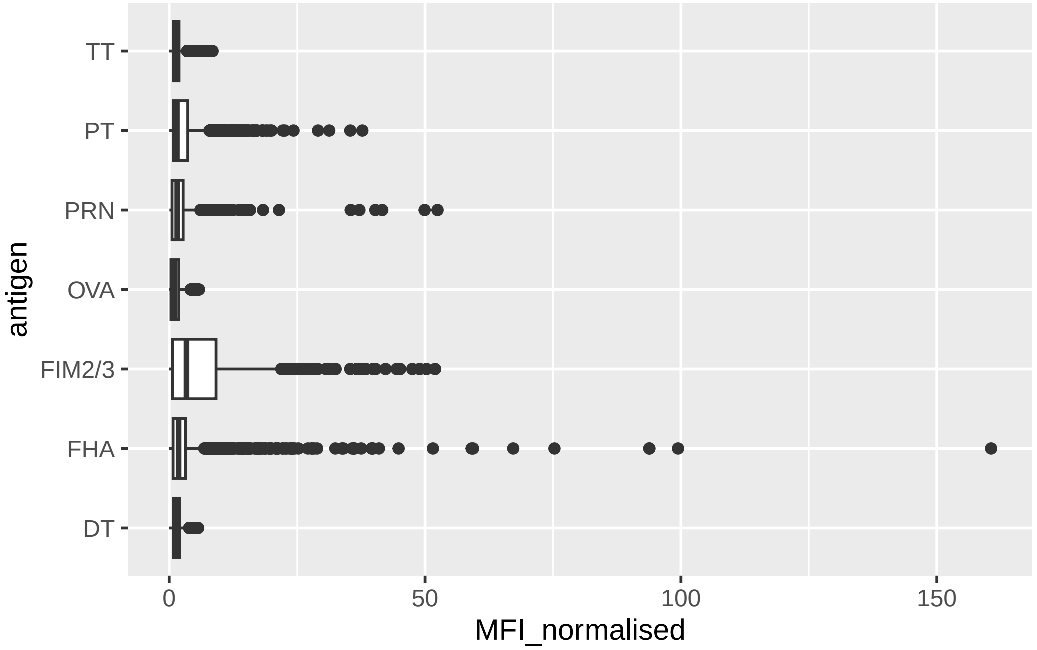
```
igg <- abdata |> filter(isotype == "IgG")

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 14394 days                           -3                             0
2 14394 days                           -3                             0
3 14394 days                           -3                             0
4 15490 days                           -3                             0
5 15490 days                           -3                             0
6 15490 days                           -3                             0
  specimen_type visit
1         Blood     1
2         Blood     1
3         Blood     1
4         Blood     1
5         Blood     1
6         Blood     1
```

Same boxplot of antigens as before

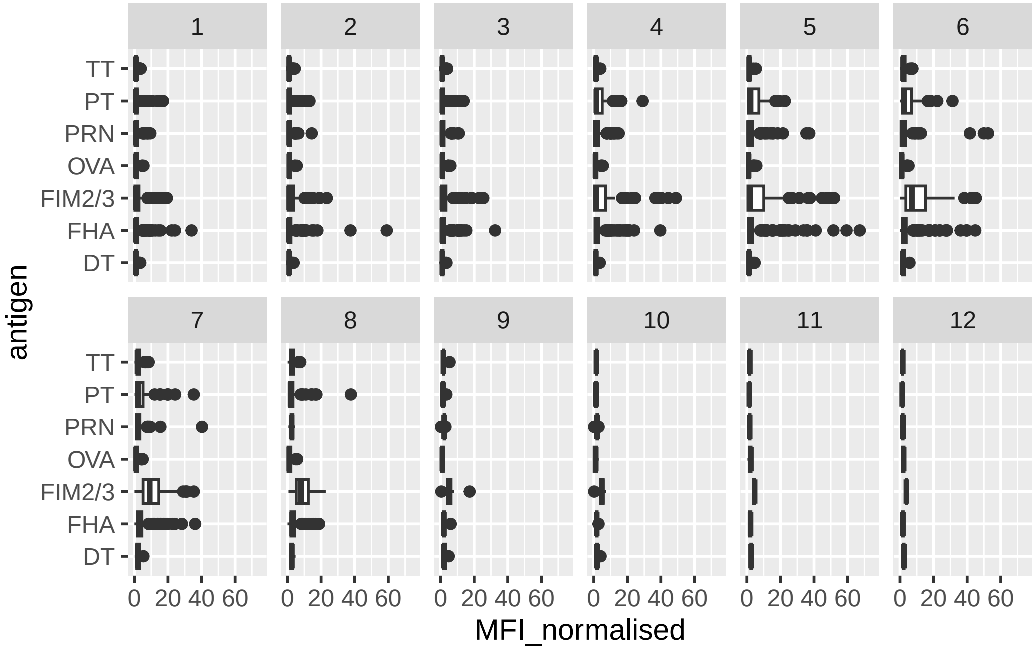
```
ggplot(igg) +
  aes(MFI_normalised, antigen) +
  geom_boxplot()
```



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()`).
```

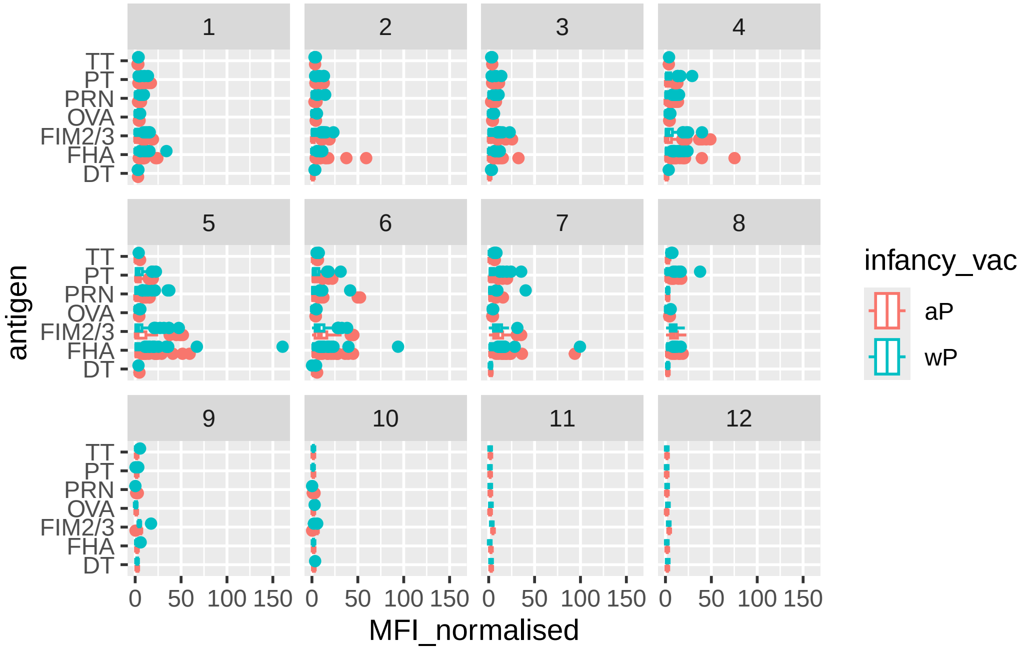


> Q14. What antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others?

PT, PRN, FIM2/3, and FHA show changes in IgG antibody titers recognizing them over time. These antigens change because these antigens are also found on the bacteria that causes Pertussis.

Look at the differences of the antigens between the aP and wP groups:

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



Focus in further in just one of these antigens - let's pick **PT** (Pertussis Toxin, one of the main toxins of the bacteria) in the **2021_dataset** again for **IgG** antibody isotypes.

```r
table(igg$dataset)
```

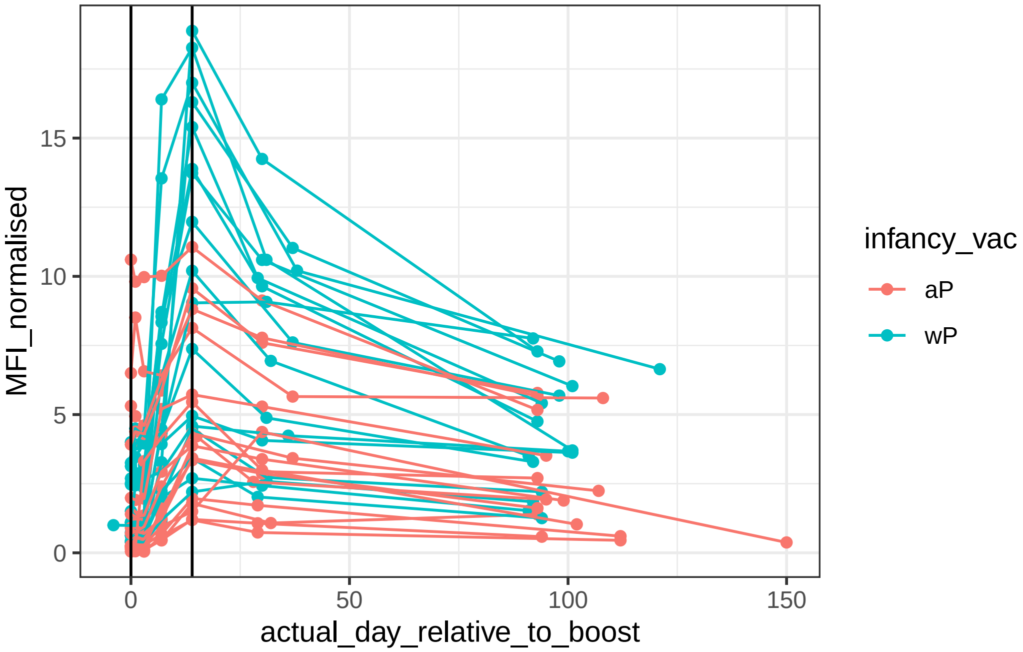
```
2020_dataset 2021_dataset 2022_dataset 2023_dataset 
        1182         1617         1456         3010 
```

```r
pt_igg <- abdata |>
  filter(isotype=="IgG", antigen=="PT", dataset=="2021_dataset")
```

```r
dim(pt_igg)
```

```
[1] 231  21
```

```r
ggplot(pt_igg) +
  aes(actual_day_relative_to_boost,
      MFI_normalised,
      col=infancy_vac,
      group = subject_id) +
  geom_point() +
  geom_line() +
  theme_bw() +
  geom_vline(xintercept = 0) +
  geom_vline(xintercept = 14)
```



On day 14, you get peak levels in both aP and wP individuals.