Class 18: Pertussis Mini Project

Lily Huynh (PID: A16929651)

2025-03-06

Table of contents

Pertussis (a.k.a.) Whooping Cough is a deadly lung infection caused by the bacteria B. Pertussis.

The CDC tracks Pertussis cases around the US. http://tinyurl.com/pertussiscdc

We can "scrape" this data using the R **datapasta** package.

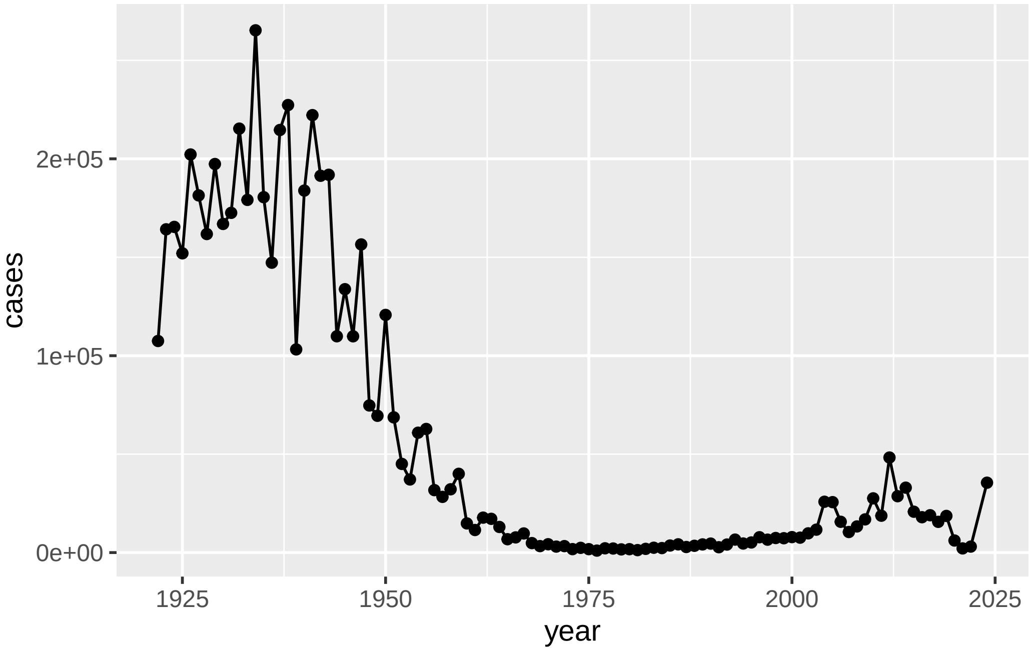
```
head(cdc)
```

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

1. Investing pertussis cases by year

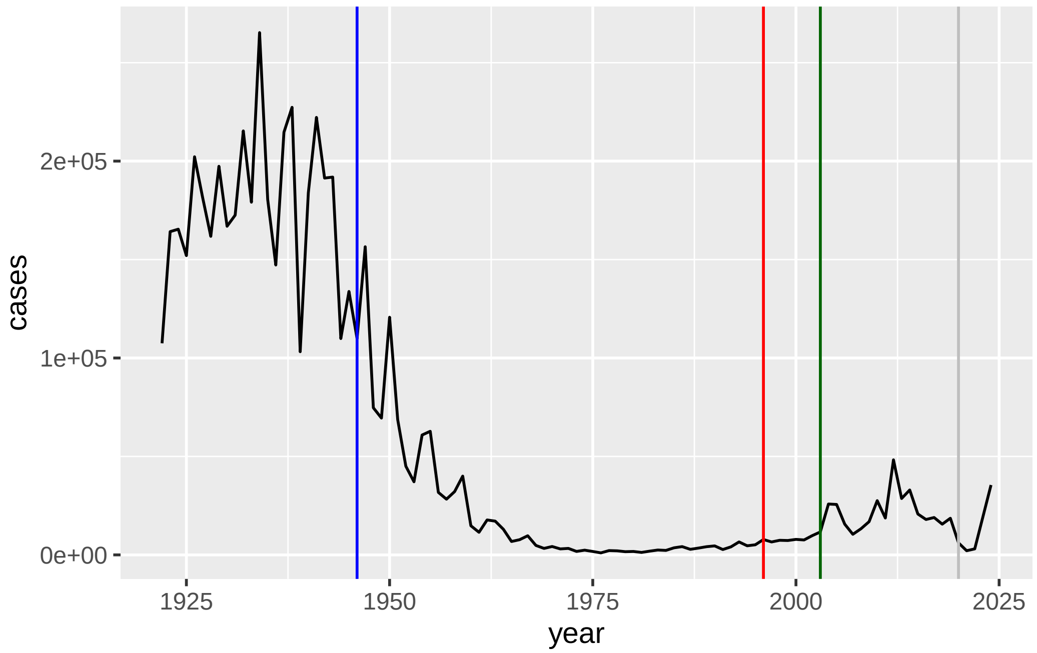
> Q1. With the help of the R "addin" package datapasta assign the CDC pertussis case number data to a data frame called cdc and use ggplot to make a plot of cases numbers over time.

```r
library(ggplot2)

# data
# aes
# geoms

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



2. A tale of two vaccines (wP&aP)

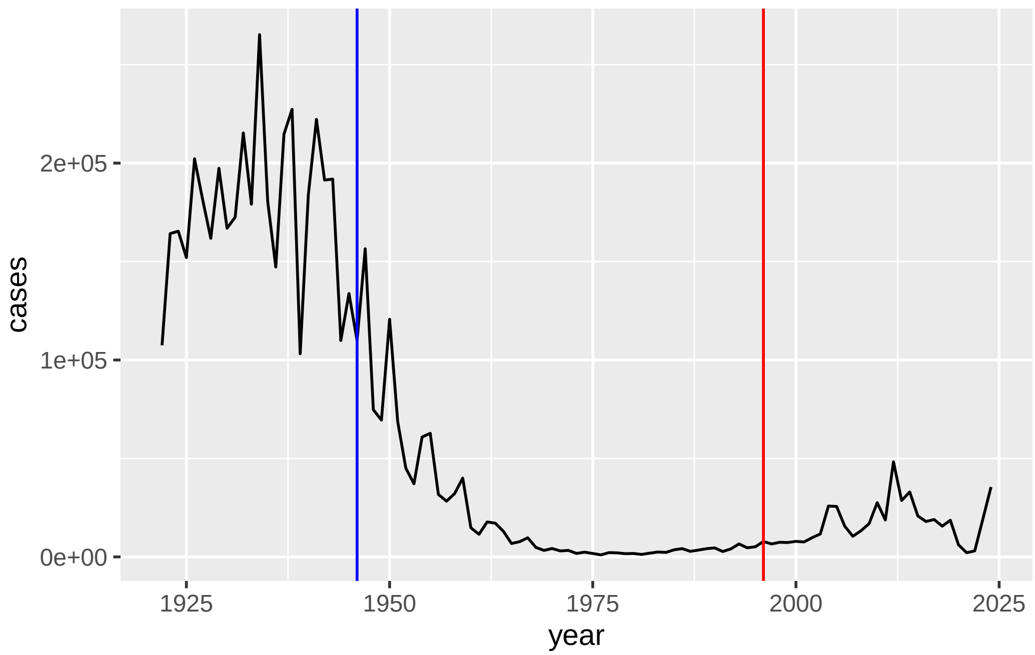
```
ggplot(cdc) +
  aes(year, cases) +
  geom_line() +
  geom_vline(xintercept = 1946, col="blue") +
  geom_vline(xintercept = 1996, col="red") +
  geom_vline(xintercept = 2020, col="grey") +
  geom_vline(xintercept = 2003, col="darkgreen")
```



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?

Ans. I notice that after the wP vaccine, the number of cases have signficantly decreased. The number of cases are close to zero. After the aP vaccine, I noticed that the number of cases has started to increase again around 2004.

```
ggplot(cdc) +
  aes(year, cases) +
  geom_line() +
  geom_vline(xintercept = 1946, col="blue") +
  geom_vline(xintercept = 1996, col="red")
```



There were high case numbers before the first wP (whole-cell) vaccine roll out in 1946 then a rapid decline in case numbers until 2004 when we have our first large-scale outbreak of pertussis again. There is also a notable COVID related dip and recent rapid rise.

Q. What is different about the immune response to infection if you had an older wP vaccine vs the newer aP vaccine?

> Q3. Describe what happened after the introduction of the aP vaccine? Do you have a possible explaination for the observed trend?
>
> Ans. After the introduction of the aP vaccine, the number of cases started to increase around 8 years after (2004). This could probably be due to a decrease in immunity over time. Since the aP doesn't contain the whole cell, the antibodies might have a hard time recognizing pertussis as the years pass. Therefore, booster shots are needed.

3. Computational Models of Immunity - Pertussis Boost (CMI-PB)

The CMI-PB project aims to address this key question: What is the difference between aP and wP individuals.

We can get all the data from this ongoing project via JSON API calls. For this we will use the **jsonlite** package. We can install with: `install.packages("jsonlite")`

```r
library(jsonlite)
```

```
Warning: package 'jsonlite' was built under R version 4.4.3
```

```r
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 individuals "subjects" are in this data set?

```r
nrow(subject)
```

```
[1] 172
```

> Q4. How many aP and wP infancy vaccinated subjects are in the dataset?
>
> Ans. There are 87 aP and 85 wP infancy vaccinated subjects in the dataset.

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

```

aP wP 
87 85 
```

> Q5. How many Male and Female subjects/patients are in the dataset?
>
> Ans. There are 112 females and 60 males in this 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
```

This is not representative of the US population but it is the biggest dataset of its type so let's see what we can learn…

> Q7. Using this approach determine (i) the average age of wP individuals, (ii) the average age of aP indivuals, and (iii) are they signficantly different?
>
> Ans. The average age of wP individuals are 35.83 years and the average age of aP individuals are 27.08 years. Since the calculated p-value is less than 0.05, they are signifcantly different.

```
library(lubridate)
```

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

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

    date, intersect, setdiff, union
```

```r
## First I need to calculate how old they are now in days

subject$age <- today() - ymd(subject$year_of_birth)

head(subject$age)
```

```
Time differences in days
[1] 14311 20886 15407 13581 12485 13581
```

```r
# Calculate the age of aP and wP

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
```

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

summary(time_length(aP$age, "years"))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  22.18   26.18   27.18   27.08   28.18   34.18 
```

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

summary(time_length(wP$age, "years"))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  22.18   32.18   34.18   35.83   39.18   57.18
```

```
# Calculate the p.value

result <- t.test(time_length(wP$age, "years"), time_length(aP$age, "years"))
result$p.value
```

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

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

Ans. See the code below

```
boost_time <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)

age_at_boost <- time_length(boost_time, "year")

age_at_boost
```

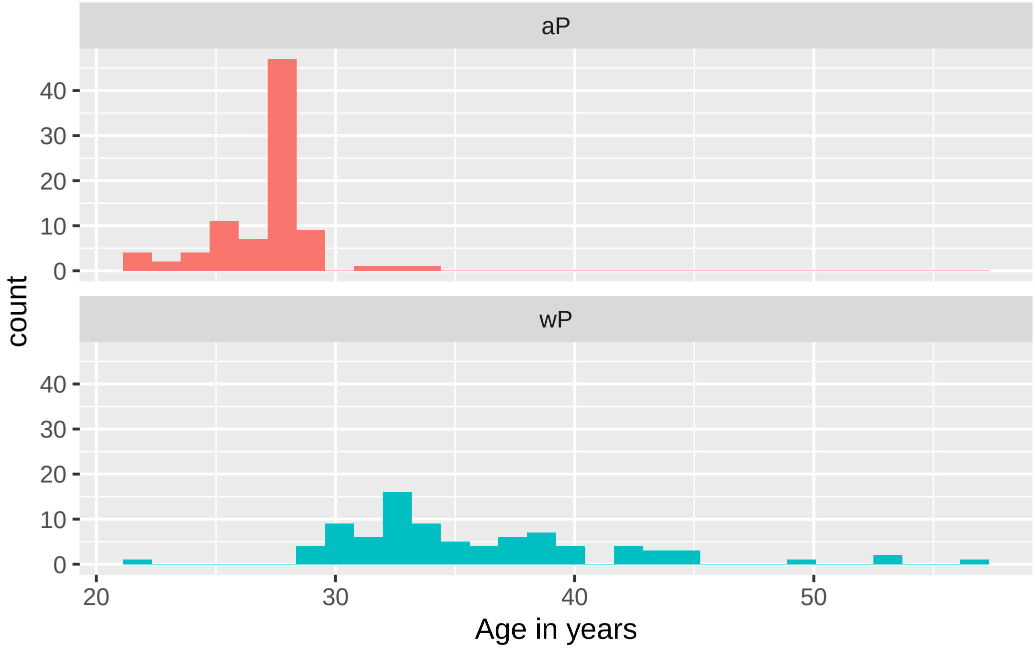
```
  [1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481 35.84942 34.14921
  [9] 20.56400 34.56263 30.65845 34.56263 19.56194 23.61944 27.61944 29.56331
 [17] 36.69815 19.65777 22.73511 35.65777 33.65914 31.65777 25.73580 24.70089
 [25] 28.70089 33.73580 19.73443 34.73511 19.73443 28.73648 27.73443 19.81109
 [33] 26.77344 33.81246 25.77413 19.81109 18.85010 19.81109 31.81109 22.81177
 [41] 31.84942 19.84942 18.85010 18.85010 19.90691 18.85010 20.90897 19.04449
 [49] 20.04381 19.90691 19.90691 19.00616 19.00616 20.04381 20.04381 20.07940
 [57] 21.08145 20.07940 20.07940 20.07940 32.26557 25.90007 23.90144 25.90007
 [65] 28.91992 42.92129 47.07461 47.07461 29.07324 21.07324 21.07324 28.15058
 [73] 24.15058 24.15058 21.14990 21.14990 31.20876 26.20671 32.20808 27.20876
 [81] 26.20671 21.20739 20.26557 22.26420 19.32375 21.32238 19.32375 19.32375
 [89] 22.41752 20.41889 21.41821 19.47707 23.47707 20.47639 21.47570 19.47707
 [97] 35.90965 28.73648 22.68309 20.83231 18.83368 18.83368 27.68241 32.68172
[105] 27.68241 25.68378 23.68241 26.73785 32.73648 24.73648 25.79603 25.79603
[113] 25.79603 31.79466 19.83299 21.91102 27.90965 24.06297 23.90965 27.12115
[121] 22.12183 23.12115 26.17933 22.17933 29.17728 29.23477 26.23682 28.29295
[129] 31.29363 26.29432 24.35044 27.35113 25.40999 32.41068 27.56194 27.41136
[137] 24.50650 22.56263 29.56057 21.69473 26.69678 31.90691 19.90691 23.90691
[145] 20.90623 31.00616 23.00616 35.00616 32.00548 32.00548 31.04449 28.12047
[153] 25.11978 26.11910 26.19302 22.19302 26.19302 23.19507 29.19370 27.32923
[161] 30.32717 24.55852 30.55715 32.55852 30.55715 22.67488 26.67488 32.67625
[169] 20.67625 31.75086 20.86516 36.06297
```

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

Ans. Based on the histograms, I think these two groups are signficantly 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`.
```



Obtain more data from CMI-PB:

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

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

```r
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
```

```r
head(ab_data)
```

```
  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
```

I now have 3 tables of data from CMI-PB: `subject`, `specimen`, and `ab_data`. I need to "join" these tables so I will have all the info I need to work with.

For this we will use the `inner_join()` function from the **dplyr** package.

> 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:

```r
library(dplyr)

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

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

```r
dim(meta)
```

```
[1] 1503   14
```

```r
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 14311 days
2 14311 days
3 14311 days
4 14311 days
5 14311 days
6 14311 days
```

```r
dim(subject)
```

```
[1] 172   9
```

```r
dim(specimen)
```

```
[1] 1503    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.

Now we can join our `ab_data` table to `meta` so we have all the info we need about antibody levels.

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

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

```r
dim(abdata)
```

```
[1] 61956    21
```

```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 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 14311 days
2 14311 days
3 14311 days
4 14311 days
5 14311 days
6 14311 days
```

> Q. How many different antibody isotypes are there in this dataset?

```r
length(abdata$isotype)
```

```
[1] 61956
```

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

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

```

  IgE   IgG  IgG1  IgG2  IgG3  IgG4 
 6698  7265 11993 12000 12000 12000 
```

> Q12. What are the different $dataset values in abdata and what do you notice about the number of rows for the most "recent" dataset?
>
> Ans. See table below for the different $dataset values in abdata. I noticed that the most recent dataset (2023) has way less rows compared to the oldest dataset (2020).

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

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

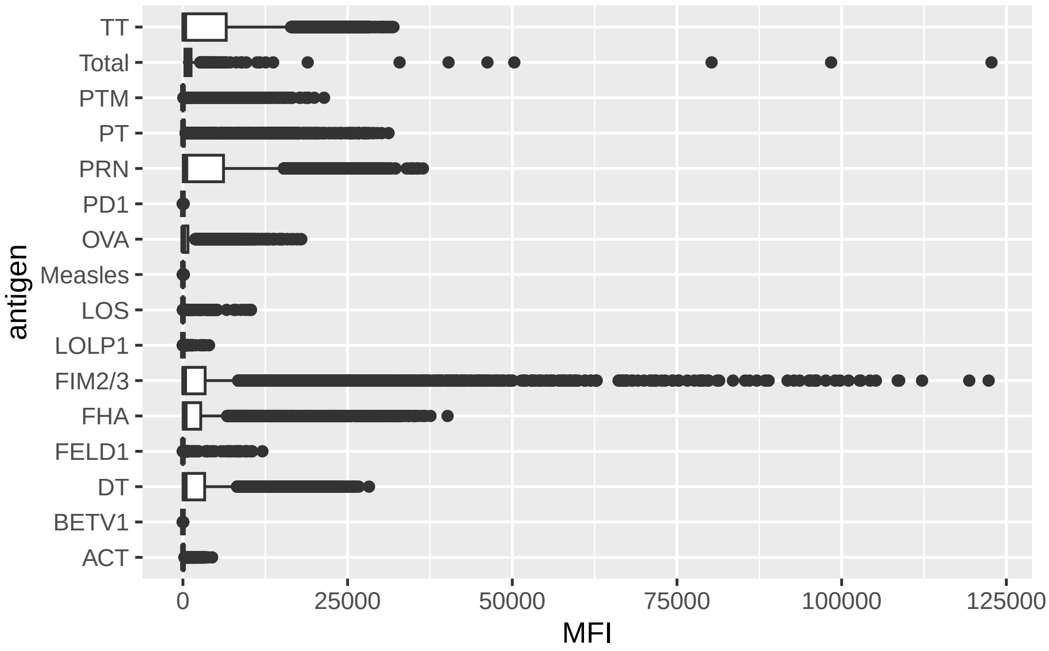
```
table(abdata$antigen)
```

```
    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
```

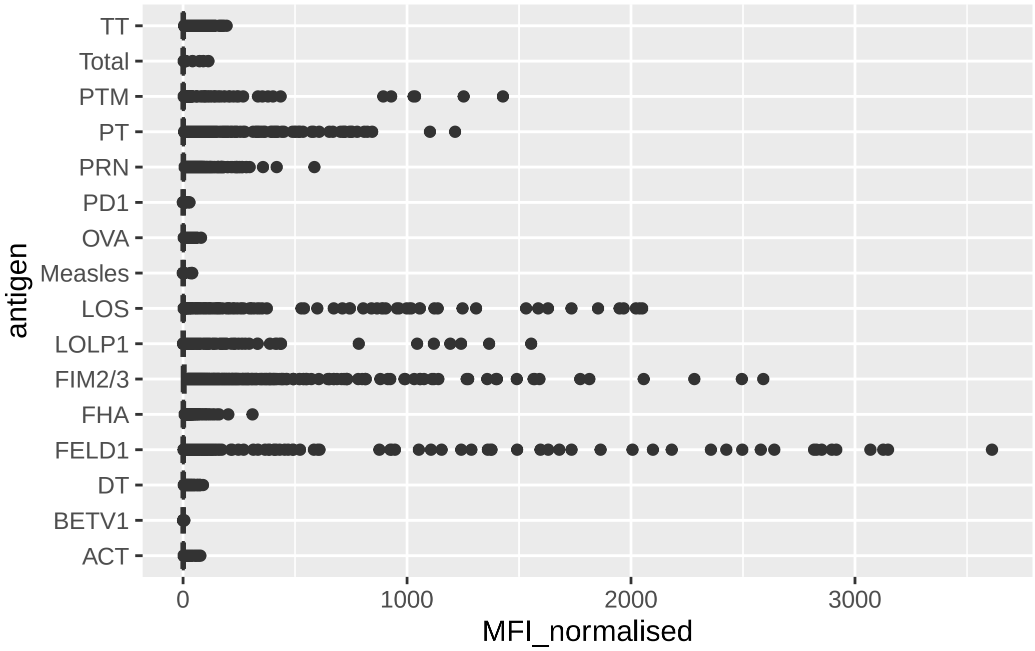
I want a plot of antigen levels across the whole data set

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

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



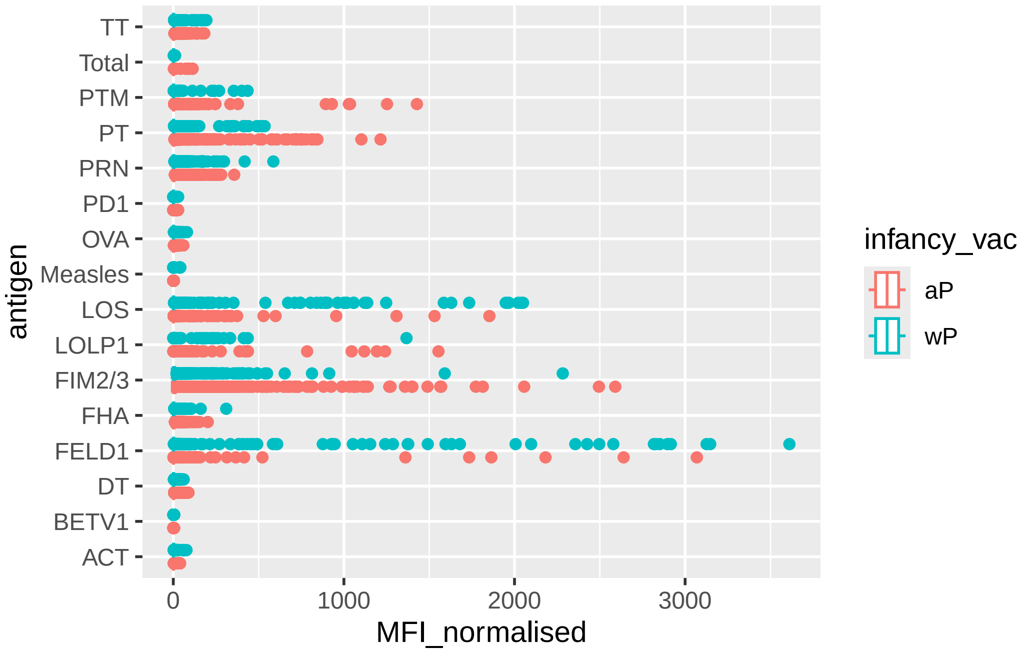
```
ggplot(abdata) +
  aes(MFI_normalised, antigen) +
  geom_boxplot()
```



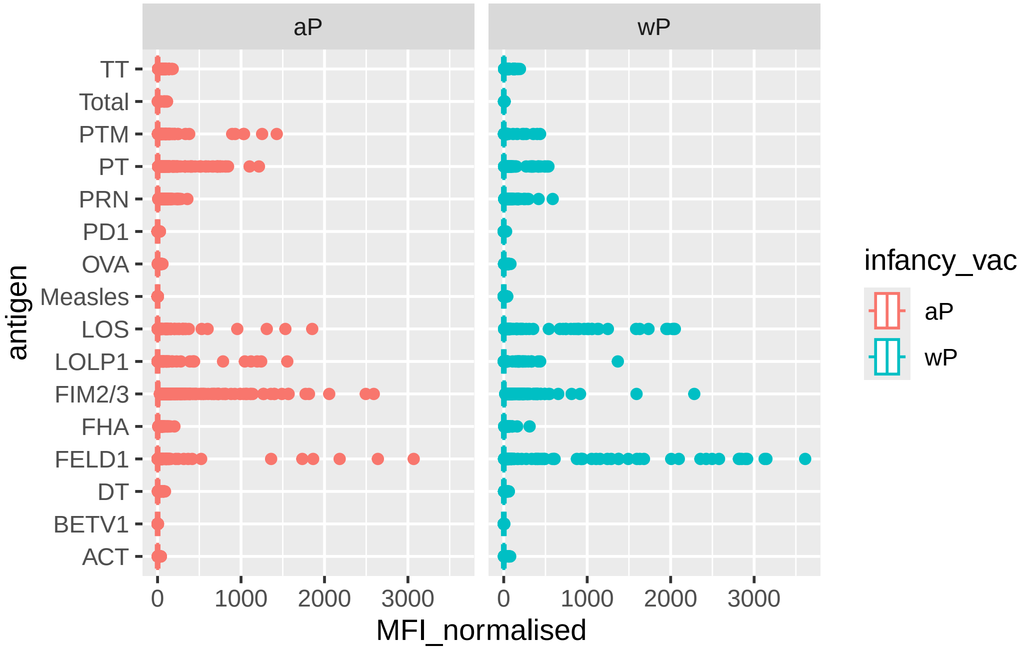
Antigens like FIM2/3, PT, and FELD1 have quite a large range of values. Others like Measles don't show much activity

> Q. Are there differences at this whole-dataset level between aP and wP?

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



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



4. Examine IgG Ab titer levels

For this, I need to select out just isotype IgG

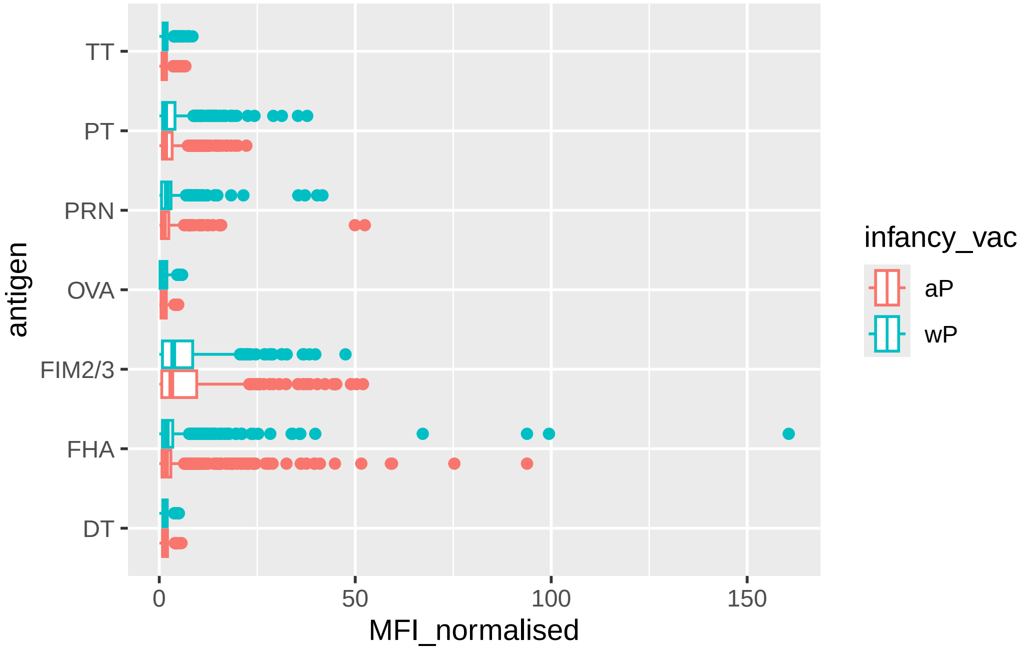
```r
igg <- abdata %>%
  filter(isotype == "IgG")
head(igg)
```

```
  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 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 14311 days
2 14311 days
3 14311 days
4 15407 days
5 15407 days
6 15407 days
```

An overview boxplot:

```r
ggplot(igg) +
  aes(MFI_normalised, antigen, col= infancy_vac) +
  geom_boxplot()
```

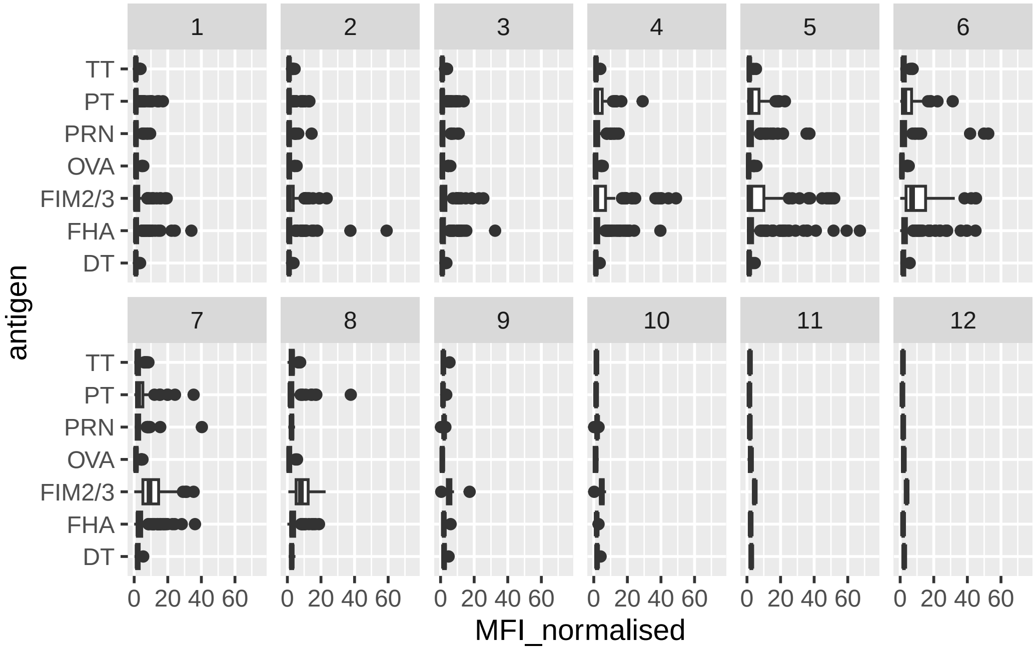


Q13. Complete the following code to make a summary boxplot of Ab titer levels (MFI) for all antigens:

See code below:

```
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()`).
```



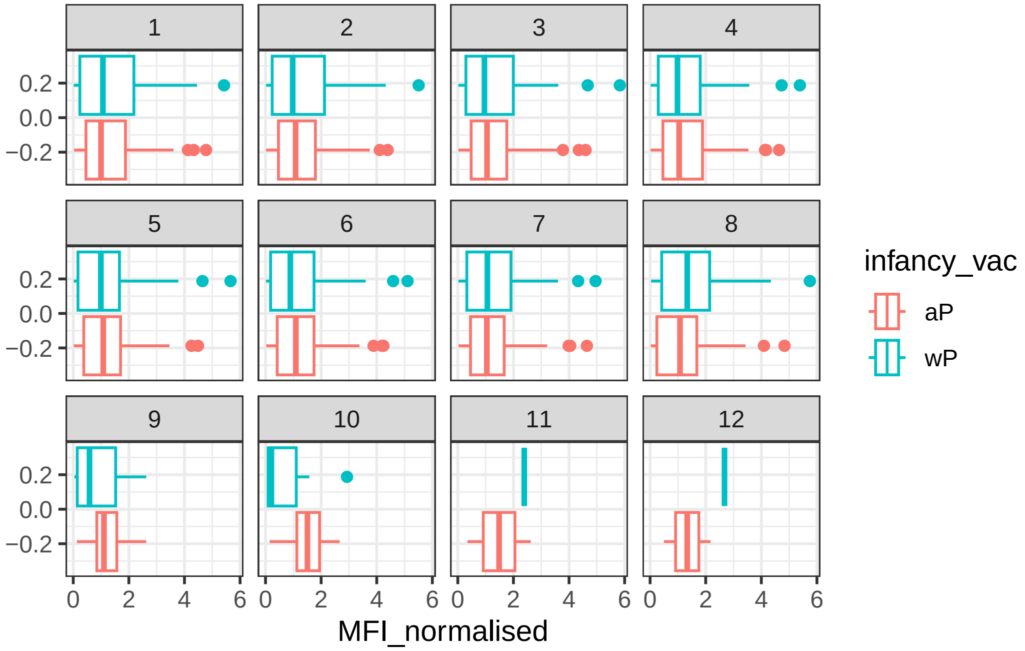
Q14. What antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others?

Ans. FIM2/3, FHA, and PRN antigens show differences in the level of IgG antibody titers recognizing them over time. These are showing difference in levels because these antigens are components of the bacteria.

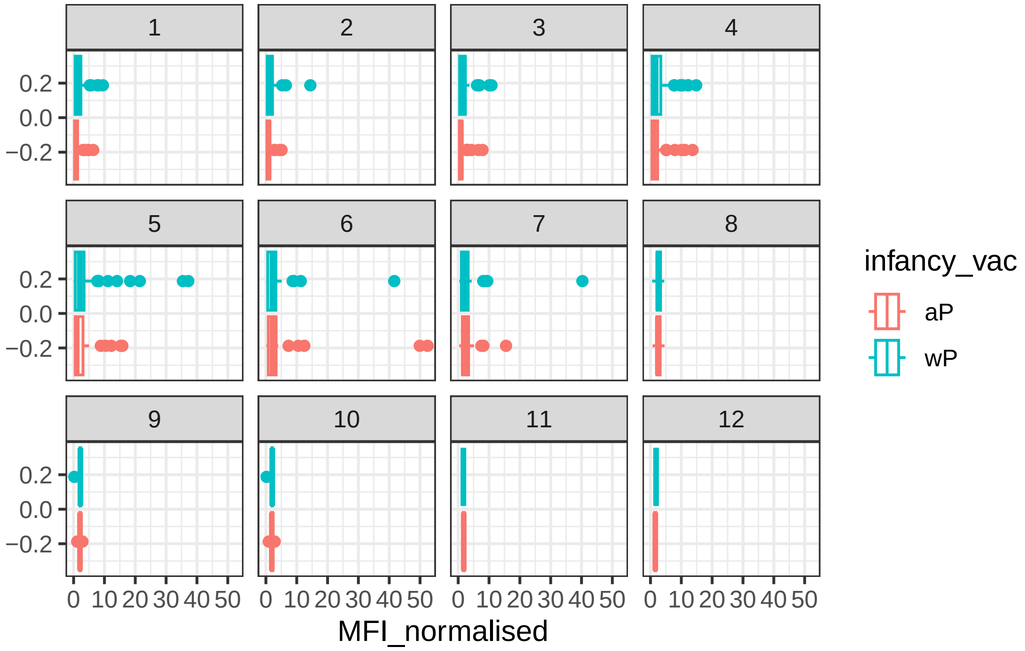
Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("OVA", that is not in our vaccines) and a clear antigen of interest ("PT", Pertussis Toxin, one of the key virulence factors produced by the bacterium B. pertussis).

Ans. Look at the code below:

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



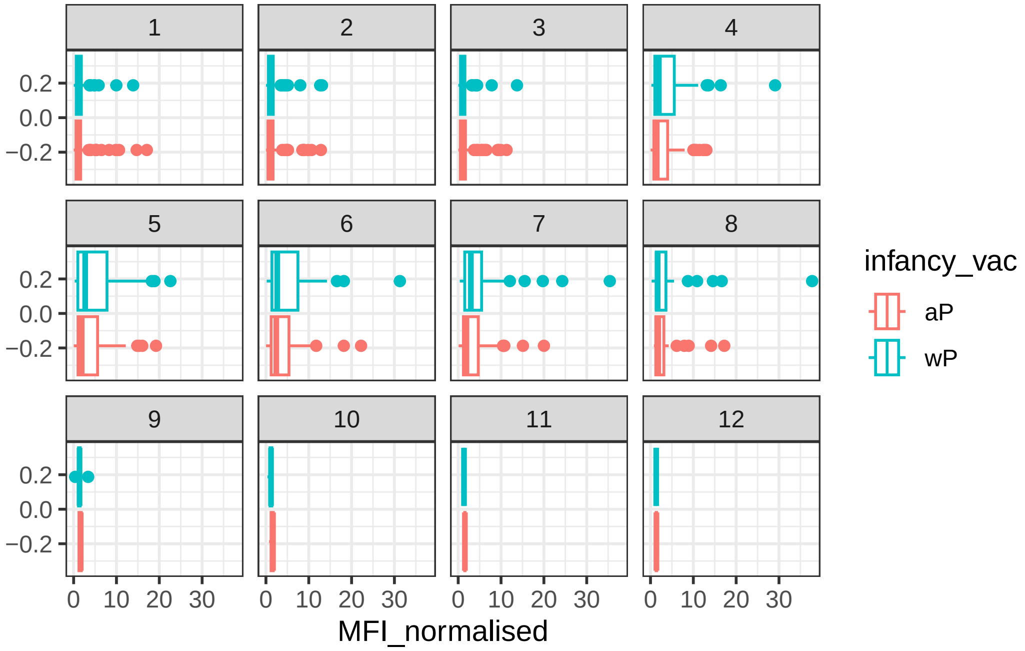
```
filter(igg, antigen=="PRN") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = TRUE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



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

Ans. I noticed that the PT and PRN antigens have higher levels compared to the OVA antigen. They rise over time, while OVA barely changes over time. The aP and wP subjects are very similar for all antigens.

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

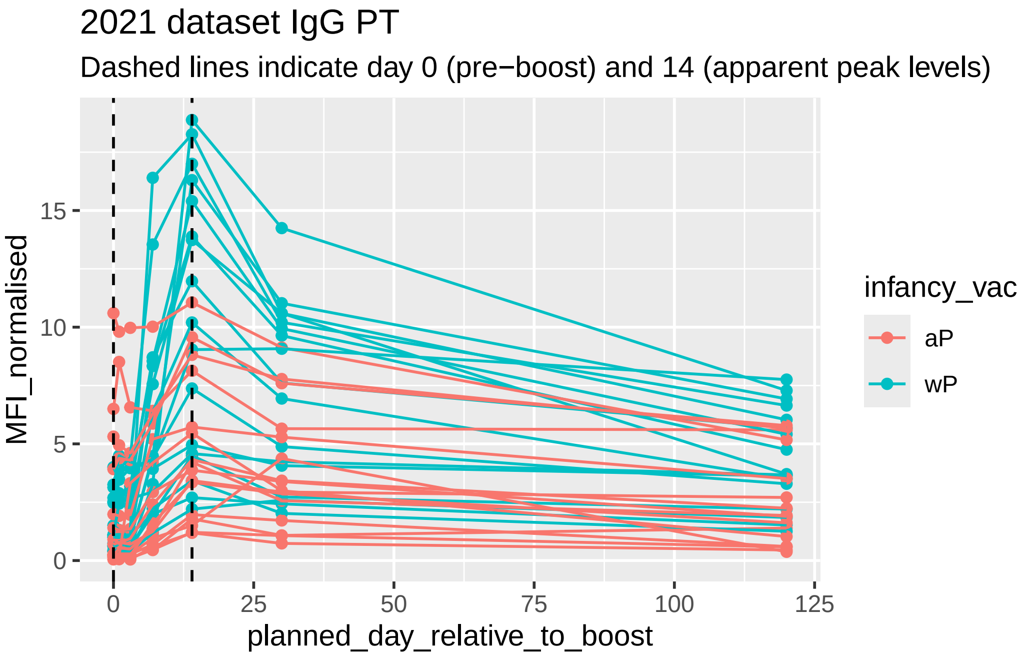


Q17. Do you see any clear difference in aP vs. wP responses?

I noticed that wP response show a bit higher values for each antigen that is in the vaccine compared to the aP. The OVA levels are pretty similar for wP and aP.

Digging in further to look at the time course of IgG isotype PT antigen levels across aP and wP individuals:

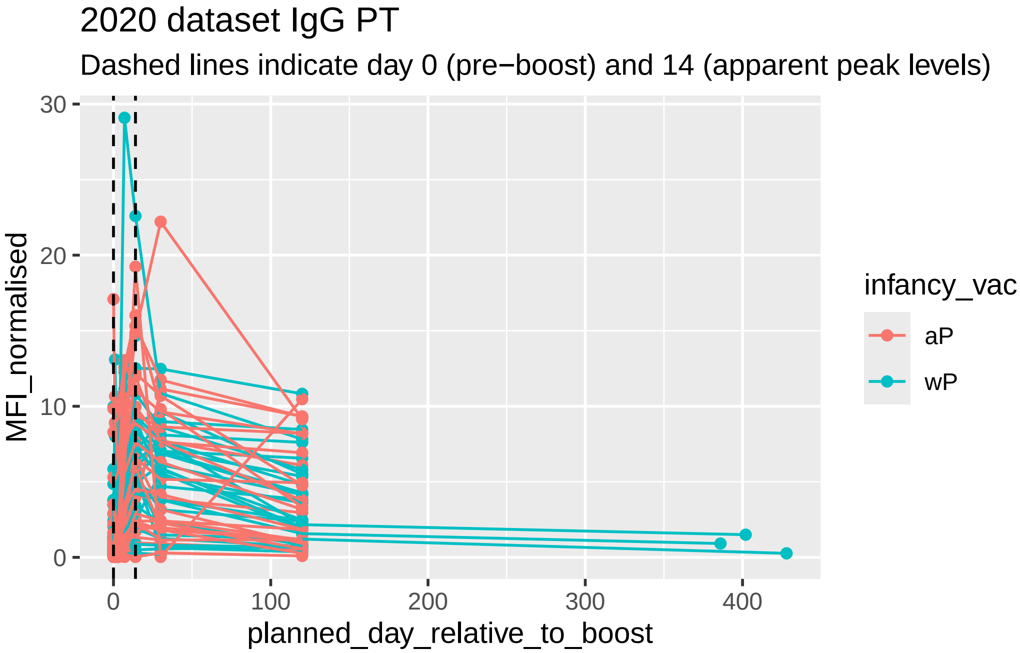
```
## Filter to include 2021 data only
abdata.21 <- abdata %>%
  filter(dataset == "2021_dataset")

## Filter to look at IgG PT data only
pt.igg <- abdata.21 %>%
  filter(isotype == "IgG",  antigen == "PT")

## Plot and color by infancy_vac (wP vs aP)
ggplot(pt.igg) +
    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)")
```



Q18. Does this trend look similar for the 2020 dataset?

Ans. This trend looks kind of similar for the 2020, but there some outliers in wP which causes the graph to look more compressed together.

```
## Filter to include 2021 data only
abdata.20 <- abdata %>%
  filter(dataset == "2020_dataset")

## Filter to look at IgG PT data only
pt.igg <- abdata.20 %>%
  filter(isotype == "IgG",  antigen == "PT")

## Plot and color by infancy_vac (wP vs aP)
ggplot(pt.igg) +
    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)")
```



5. Obtaining CMI-PB RNASeq data

```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.7"

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

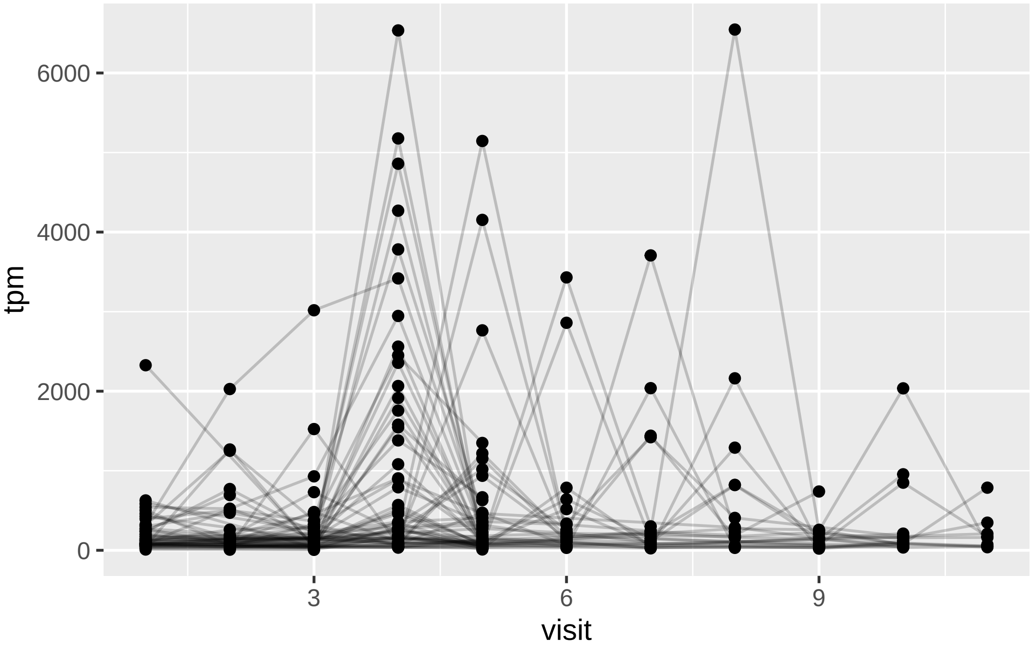
```
#meta <- inner_join(specimen, subject)
ssrna <- inner_join(rna, meta)
```

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

Q19. Make a plot of the time course of gene expression for IGHG1 gene (i.e. a plot of visit vs. tpm).

Ans. Look at code below:

```
ggplot(ssrna) +
  aes(visit, tpm, group=subject_id) +
  geom_point() +
  geom_line(alpha=0.2)
```



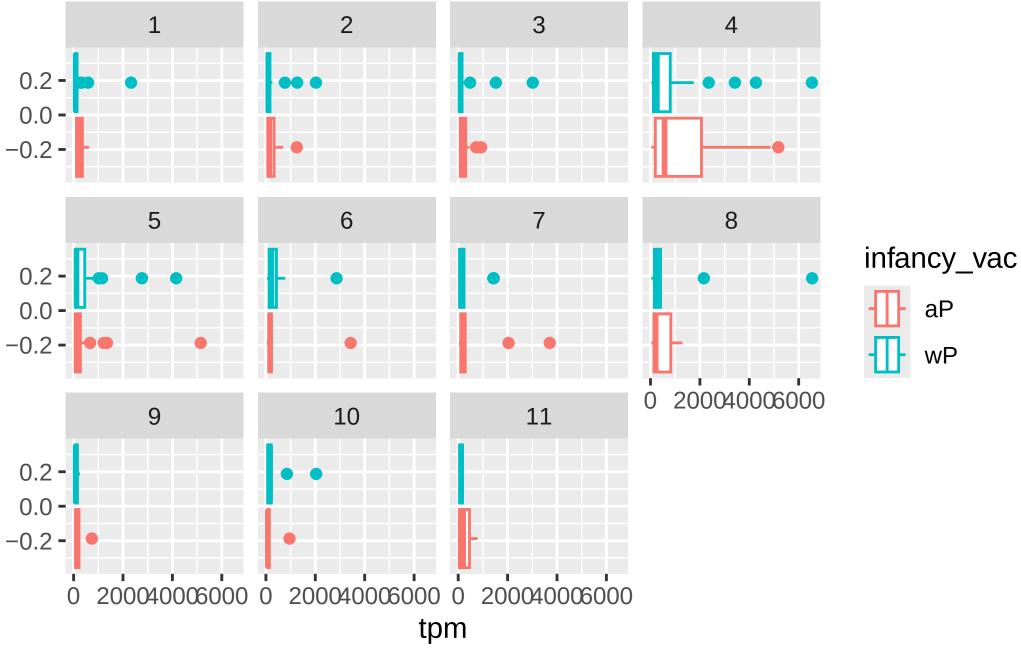
Q20.: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

Ans. The maximum levels occur on visit 4 and 8.

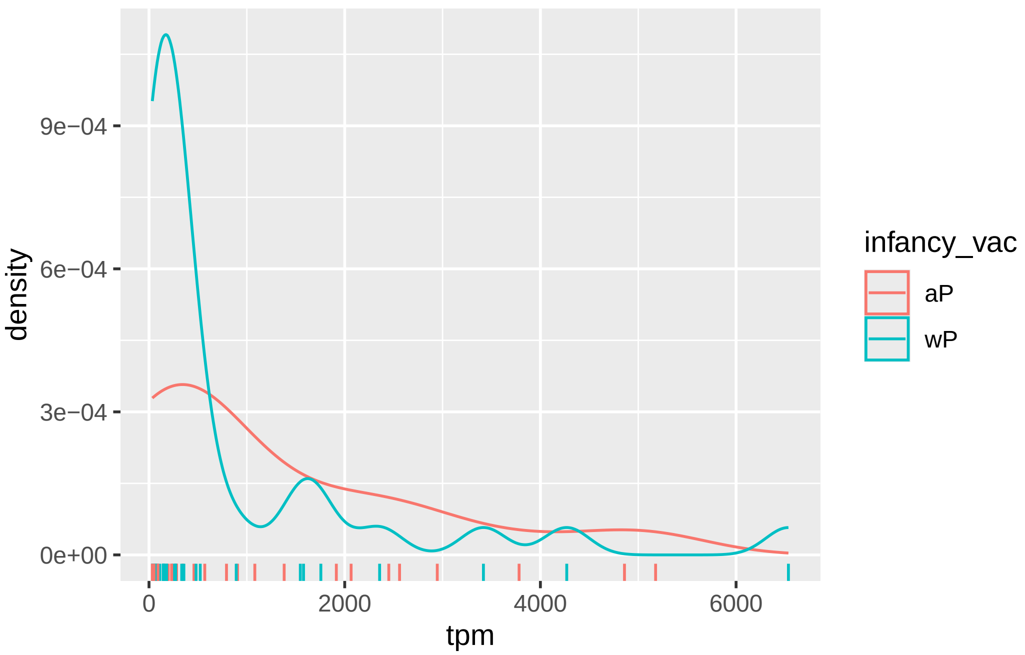
Q21. Does this pattern in time match the trend of antibody titer data? If not, why not?

Ans. The pattern in time doesn't completely match the trend of antibody titer data. The peak in antibody titer data occurs around visit 5 and then slolwy decreases. However, the pattern for time has two peaks before it decreases.

```r
ggplot(ssrna) +
  aes(tpm, col=infancy_vac) +
  geom_boxplot() +
  facet_wrap(vars(visit))
```



```r
ssrna %>%
  filter(visit==4) %>%
  ggplot() +
    aes(tpm, col=infancy_vac) + geom_density() +
    geom_rug()
```



6. Working with larger datasets [OPTIONAL]

THE DATA ISN'T DOWNLOADING because the website doesn't exist (can't do the optional)