Pertussis

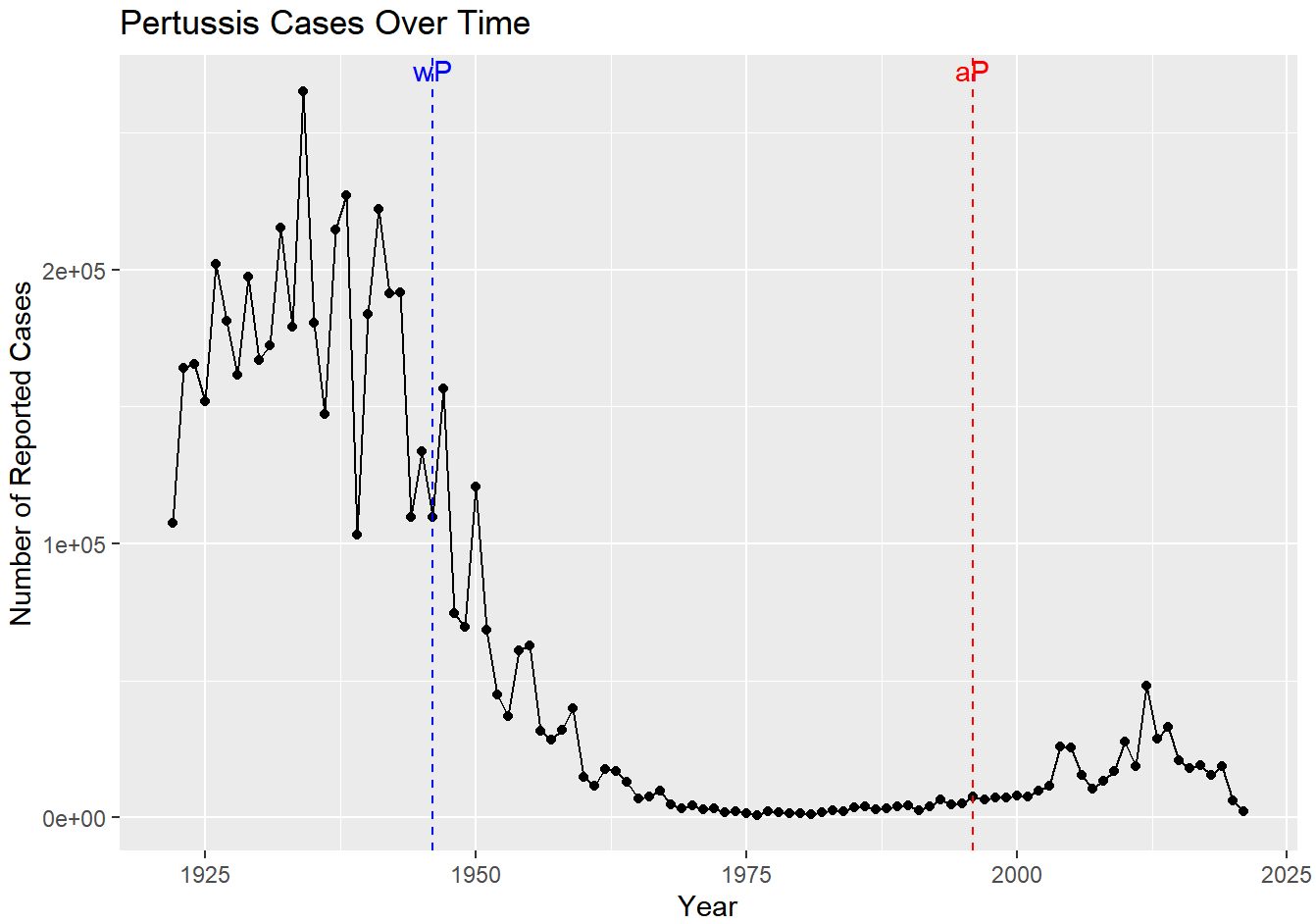
AUTHOR
Uwaysah

```
#install.packages("datapasta")
```

```
cdc <- read.csv("/Users/Uwaysah/Desktop/Pertussis/cdcdata.csv")
head(cdc)
```

```
  Year Number.of.Reported.Cases
1 1922                 107,473
2 1923                 164,191
3 1924                 165,418
4 1925                 152,003
5 1926                 202,210
6 1927                 181,411
```

```
library(ggplot2)

cdc$Number.of.Reported.Cases <- as.numeric(gsub(",", "", cdc$Number.of.Reported.Cases))

ggplot(cdc, aes(Year, Number.of.Reported.Cases)) +
  geom_point() +  # Add points
  geom_line() +   # Add lines connecting points
  labs(x = "Year", y = "Number of Reported Cases", title = "Pertussis Cases Over Time") +
  scale_y_continuous() +  # Automatically format y-axis labels with commas
  geom_vline(xintercept = 1946, linetype = "dashed", color = "blue") +  # Vertical line for wP va
  geom_vline(xintercept = 1996, linetype = "dashed", color = "red") +   # Vertical line for switcl
  annotate("text", x = 1946, y = max(cdc$Number.of.Reported.Cases), label = "wP", vjust = -0.5, c
  annotate("text", x = 1996, y = max(cdc$Number.of.Reported.Cases), label = "aP", vjust = -0.5, c
```



#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? #There was a significant decrease in the number of reported cases as soon as the wp vaccine was introduced until the ap vaccine was introduced, after which cases started to rise again.

#Q3. Describe what happened after the introduction of the aP vaccine? Do you have a possible explanation for the observed trend? #Cases started to rise again and this may be because the ap vaccine was not as effective in eliminating Pertussis in the population. The bacteria could have evolved to become more resistant to the ap vaccine after such long exposure to the wp vaccine.

```
library(jsonlite)
subject <- read_json("https://www.cmi-pb.org/api/subject", simplifyVector = TRUE)
head(subject, 3)
```

```
  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
  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
```

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

```
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$biological_sex, subject$race)
```

```
        American Indian/Alaska Native Asian Black or African American
  Female                             0    21                         2
  Male                               1    11                         0

        More Than One Race Native Hawaiian or Other Pacific Islander
  Female                  9                                        1
  Male                    2                                        1

        Unknown or Not Reported White
  Female                      11    35
  Male                         4    20
```

```
#install.packages("lubridate")
```

```
library(lubridate)
```

```
Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

```
today()
```

```
[1] "2024-03-18"
```

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

```
Time difference of 8843 days
```

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

```
[1] 24.21081
```

#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?

```
subject$age <- today() - ymd(subject$year_of_birth)
library(dplyr)
```

```
Attaching package: 'dplyr'

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

    filter, lag

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

    intersect, setdiff, setequal, union
```

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

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

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     21      26      26      26      27      30 
```

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

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     28      31      36      37      39      56 
```

#yes they are significantly different because the average age for ap individuals is 26 while the average age for wp individuals is 37.

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

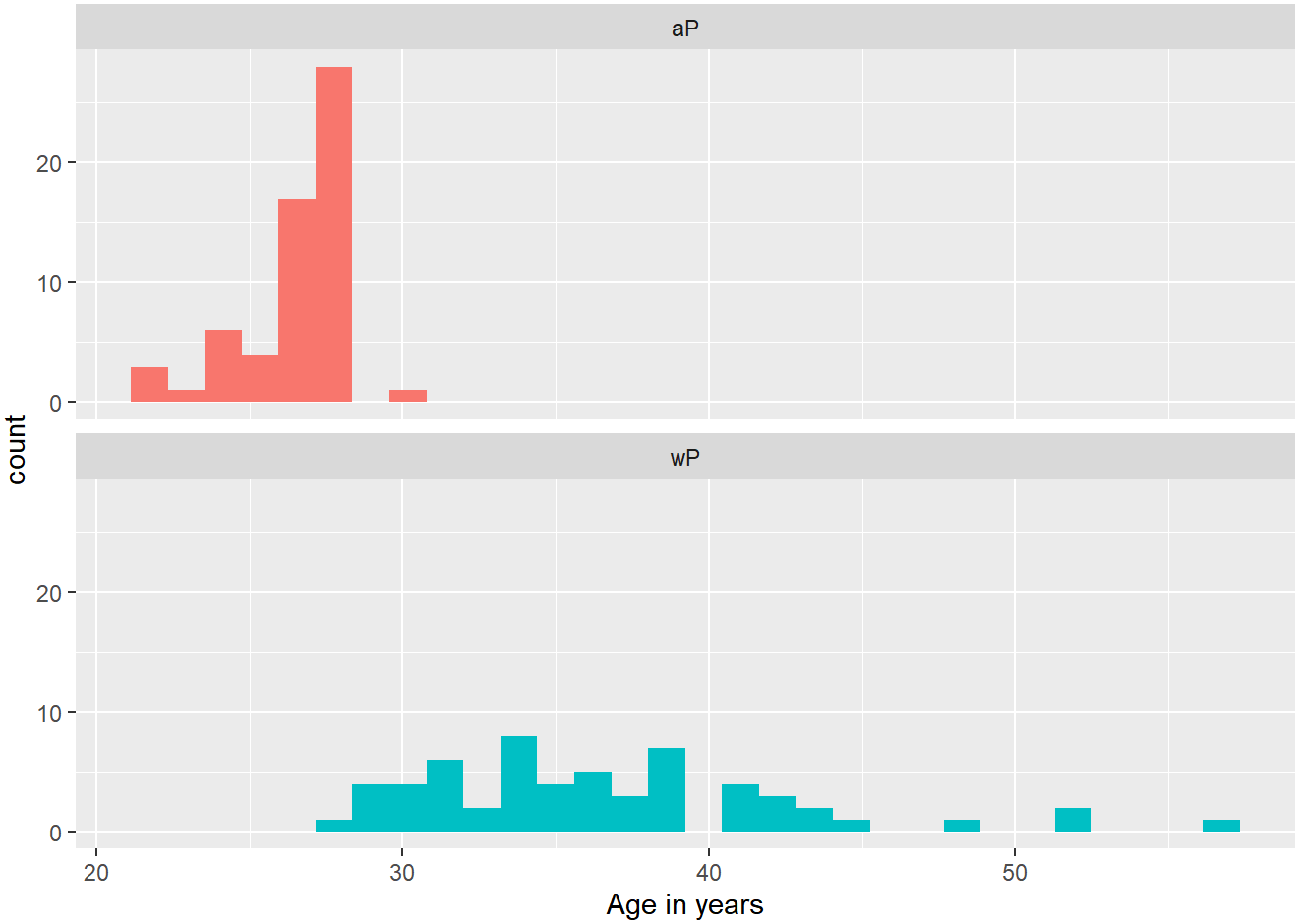
```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

```
[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 very different because their averages are far from each other and their distributions look vastly different, with the ap distribution being skewed to the left and the wp distribution being very slightly skewed to the right.

```
specimen <- read_json("https://www.cmi-pb.org/api/specimen", simplifyVector = TRUE)
titer <- read_json("https://www.cmi-pb.org/api/v4/plasma_ab_titer", simplifyVector = TRUE)
```

#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(specimen, subject)
```

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

```
dim(meta)
```

```
[1] 939  14
```

```
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 13956 days
2 13956 days
3 13956 days
4 13956 days
5 13956 days
6 13956 days
```

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

```
abdata <- inner_join(meta, titer, by = "specimen_id")
dim(abdata)
```

```
[1] 41775    21
```

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

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

```
 IgE  IgG IgG1 IgG2 IgG3 IgG4
6698 3233 7961 7961 7961 7961
```

#Q12. 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
```

#There are a lot less rows for the most recent dataset.

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

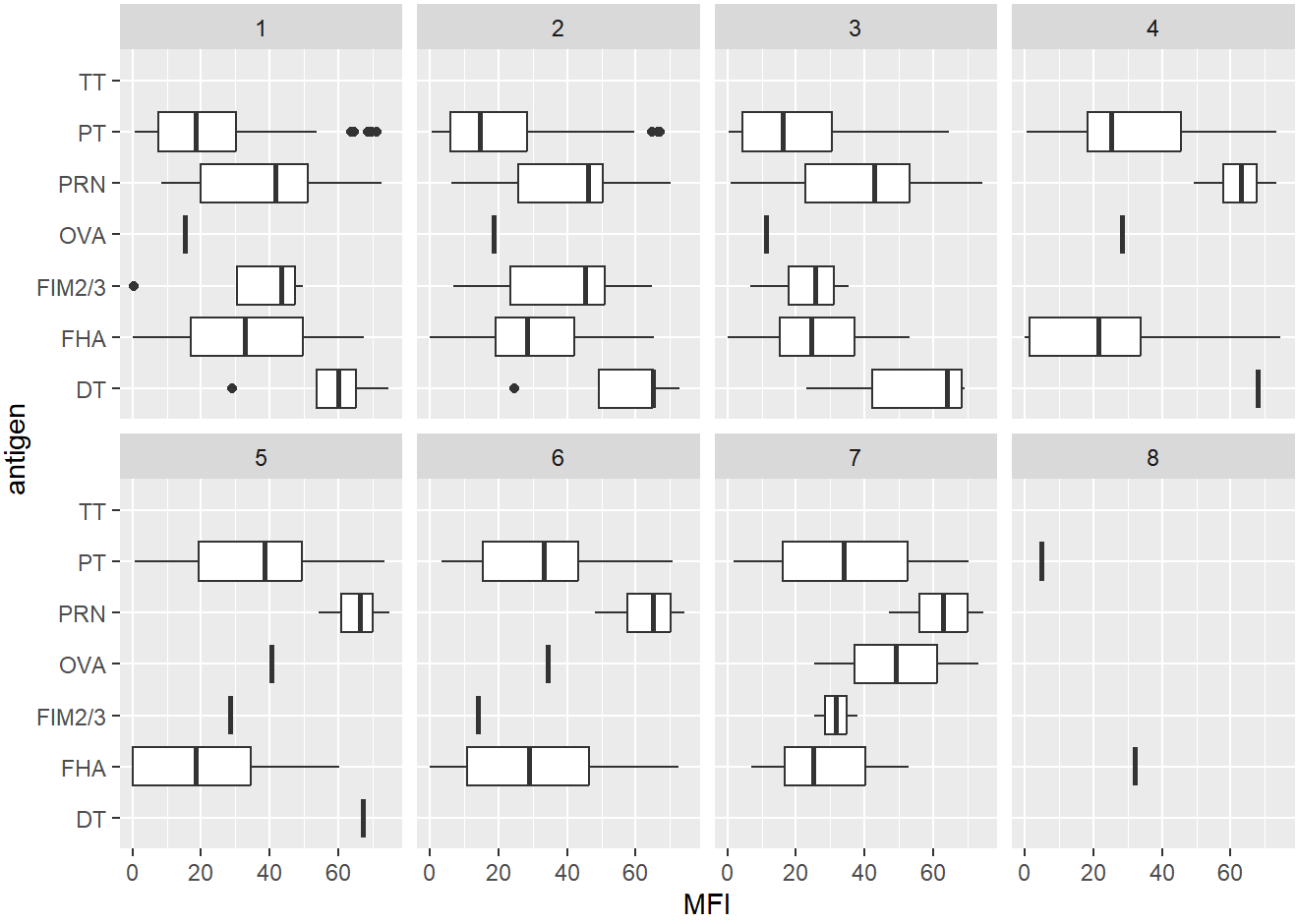
```
  specimen_id subject_id actual_day_relative_to_boost
1           1          1                           -3
2           1          1                           -3
3           1          1                           -3
4           2          1                            1
5           2          1                            1
6           2          1                            1
  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                             1         Blood     2          wP         Female
5                             1         Blood     2          wP         Female
6                             1         Blood     2          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 isotype is_antigen_specific antigen        MFI MFI_normalised
1 13956 days     IgG                TRUE      PT   68.56614       3.736992
2 13956 days     IgG                TRUE     PRN  332.12718       2.602350
3 13956 days     IgG                TRUE     FHA 1887.12263      34.050956
4 13956 days     IgG                TRUE      PT   41.38442       2.255534
5 13956 days     IgG                TRUE     PRN  174.89761       1.370393
6 13956 days     IgG                TRUE     FHA  246.00957       4.438960
   unit lower_limit_of_detection
```

```
1 IU/ML                0.530000
2 IU/ML                6.205949
3 IU/ML                4.679535
4 IU/ML                0.530000
5 IU/ML                6.205949
6 IU/ML                4.679535
```

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

```
ggplot(igg) +
  aes(x = MFI, y = antigen) +  # Specify aesthetics: MFI on x-axis, antigen on y-axis
  geom_boxplot() +              # Add boxplot
  xlim(0, 75) +                 # Set x-axis limits
  facet_wrap(vars(visit), nrow = 2)  # Facet by visit variable with 2 rows
```

```
Warning: Removed 2514 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```

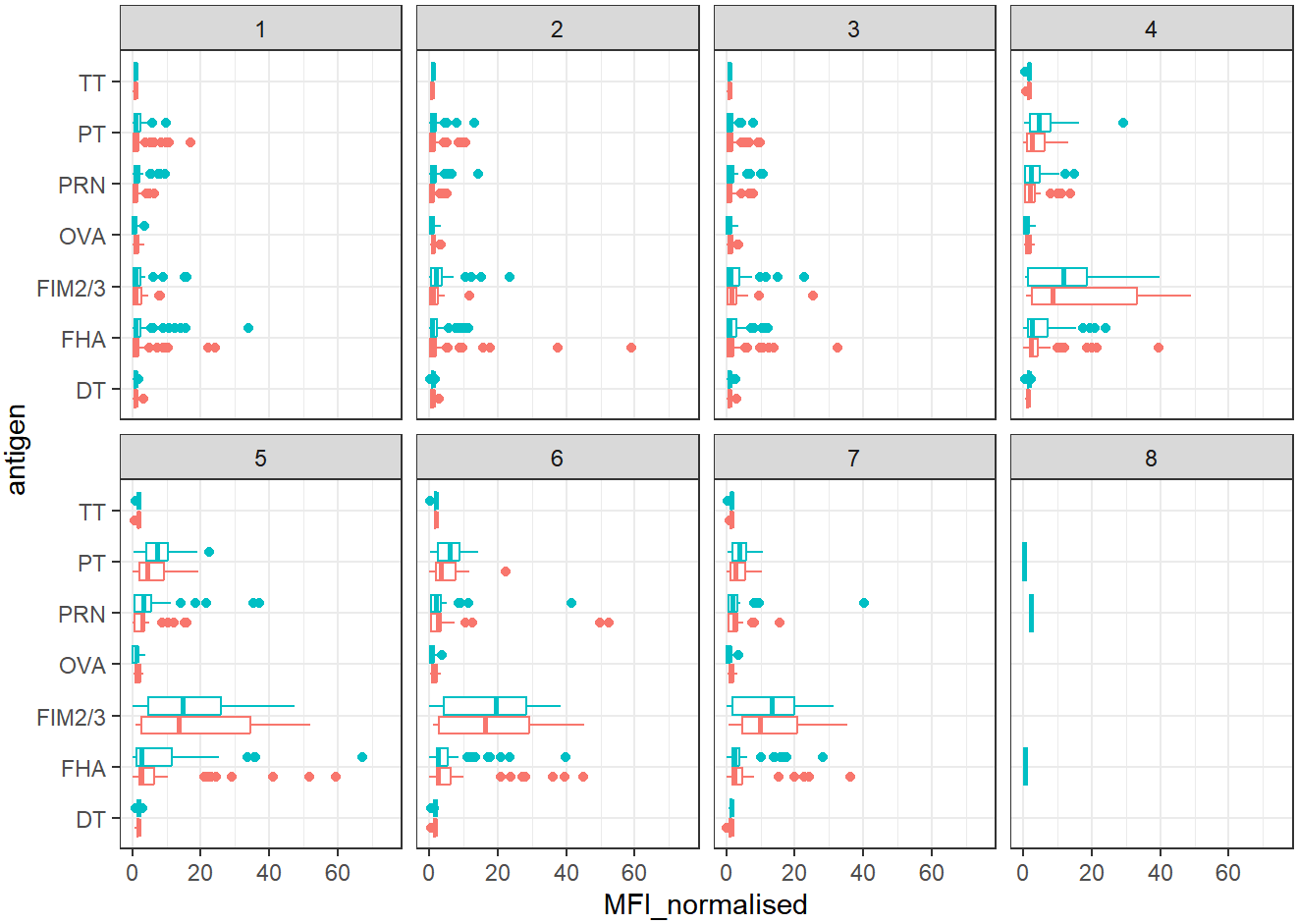


#Q14. What antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others?

#Toxoplasma gondii Antigens show differences in the level of IgG antibody titers recognizing them over time because it has developed resistence to antibodies.

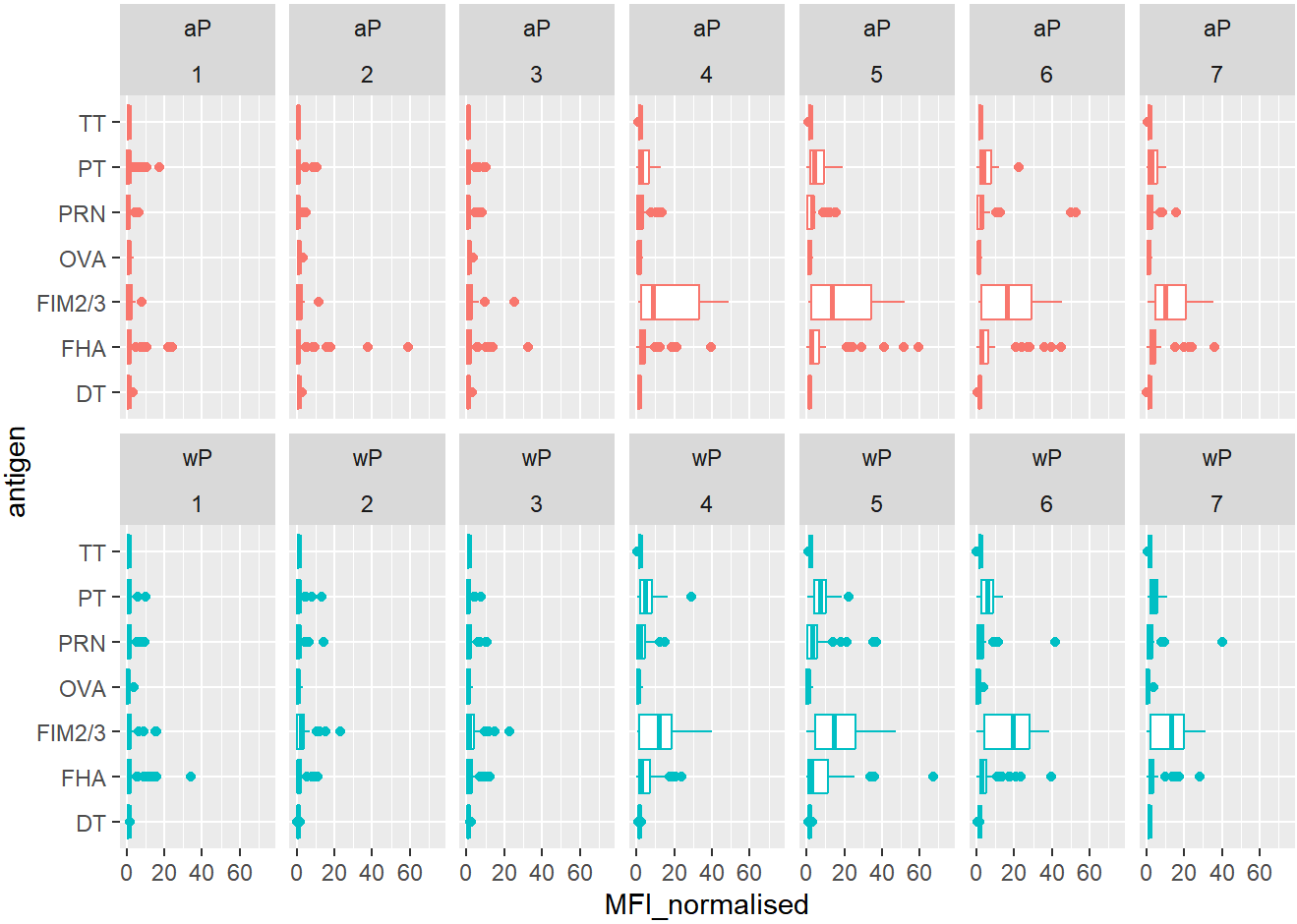
```
ggplot(igg) +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  xlim(0,75) +
  theme_bw()
```

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



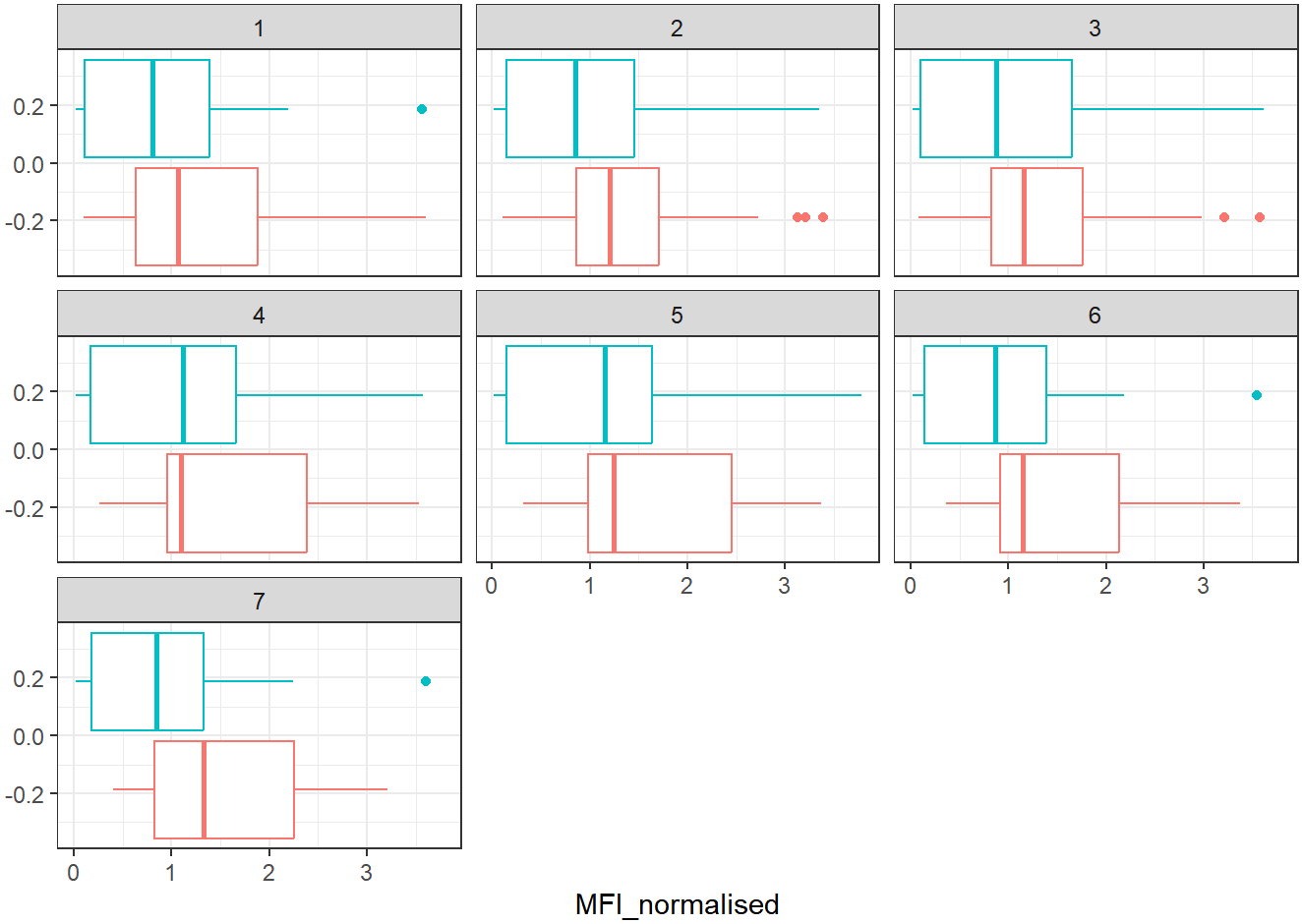

```
igg %>% filter(visit != 8) %>%
ggplot() +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  xlim(0,75) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```

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

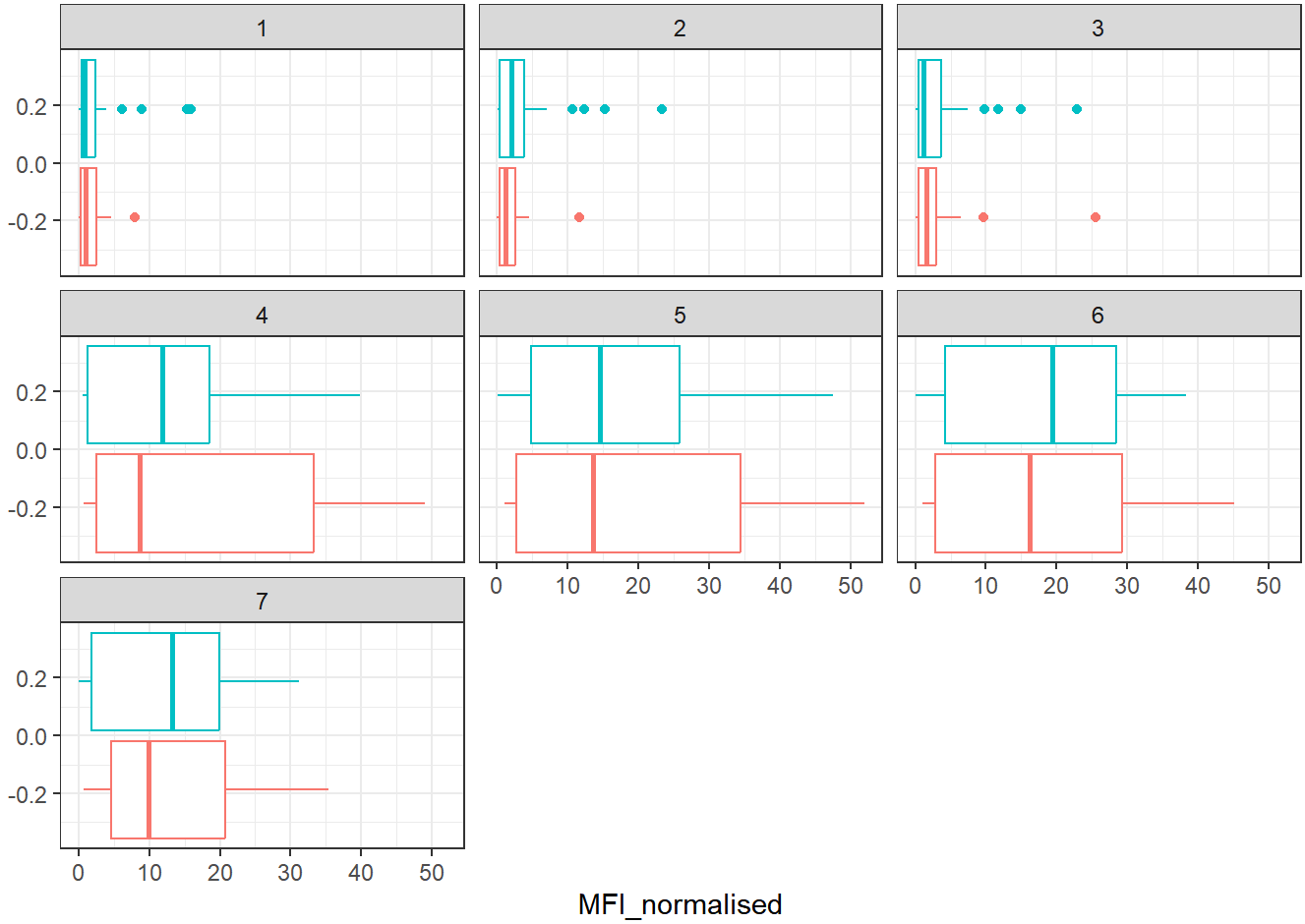


#Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("OVA", that is not in our vaccines) and a clear antigen of interest ("PT", Pertussis Toxin, one of the key virulence factors produced by the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

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




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



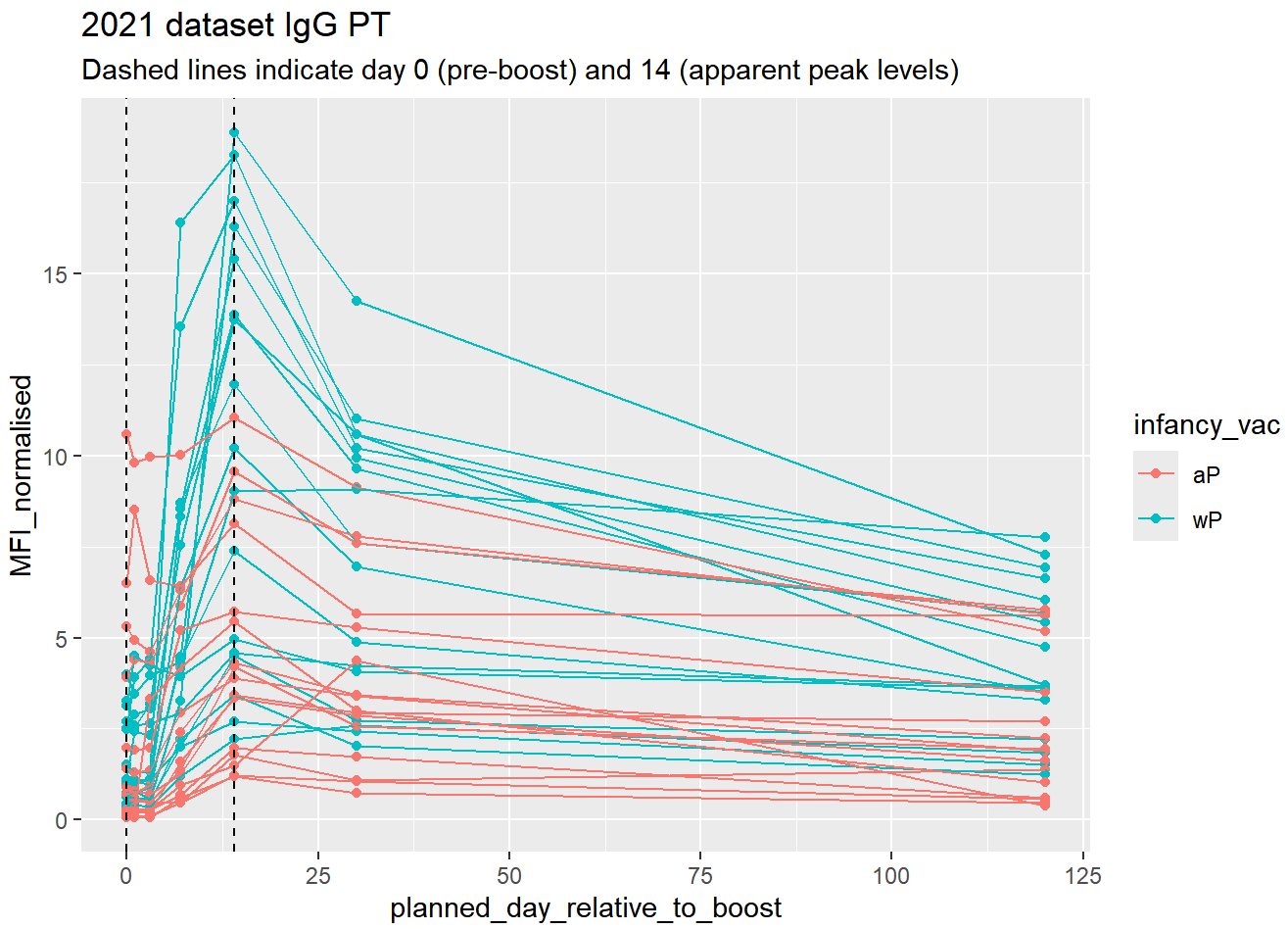
#Q16. What do you notice about these two antigens time courses and the PT data in particular?

#PT levels rise and exceed OVA levels over time. They peak at visit 6 and decline after. The trend seems the same for wp and ap participants.

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

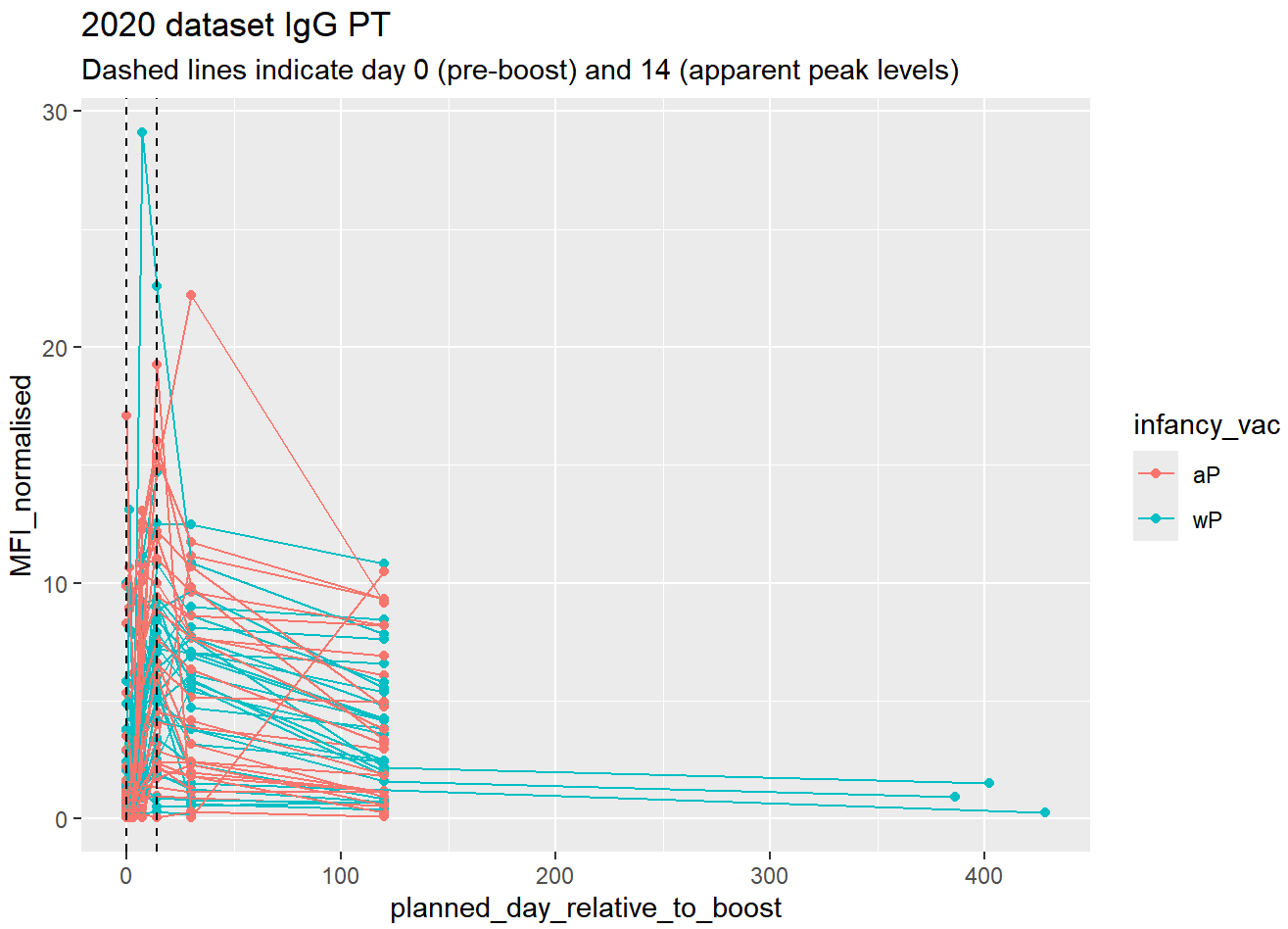
#I do not see a clear difference in ap vs wp responses.

```
abdata.21 <- abdata %>% 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)")
```



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

```
abdata.20 <- abdata %>% filter(dataset == "2020_dataset")

abdata.20 %>%
  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)")
```



#The trend does not look similar for the 2020 dataset.

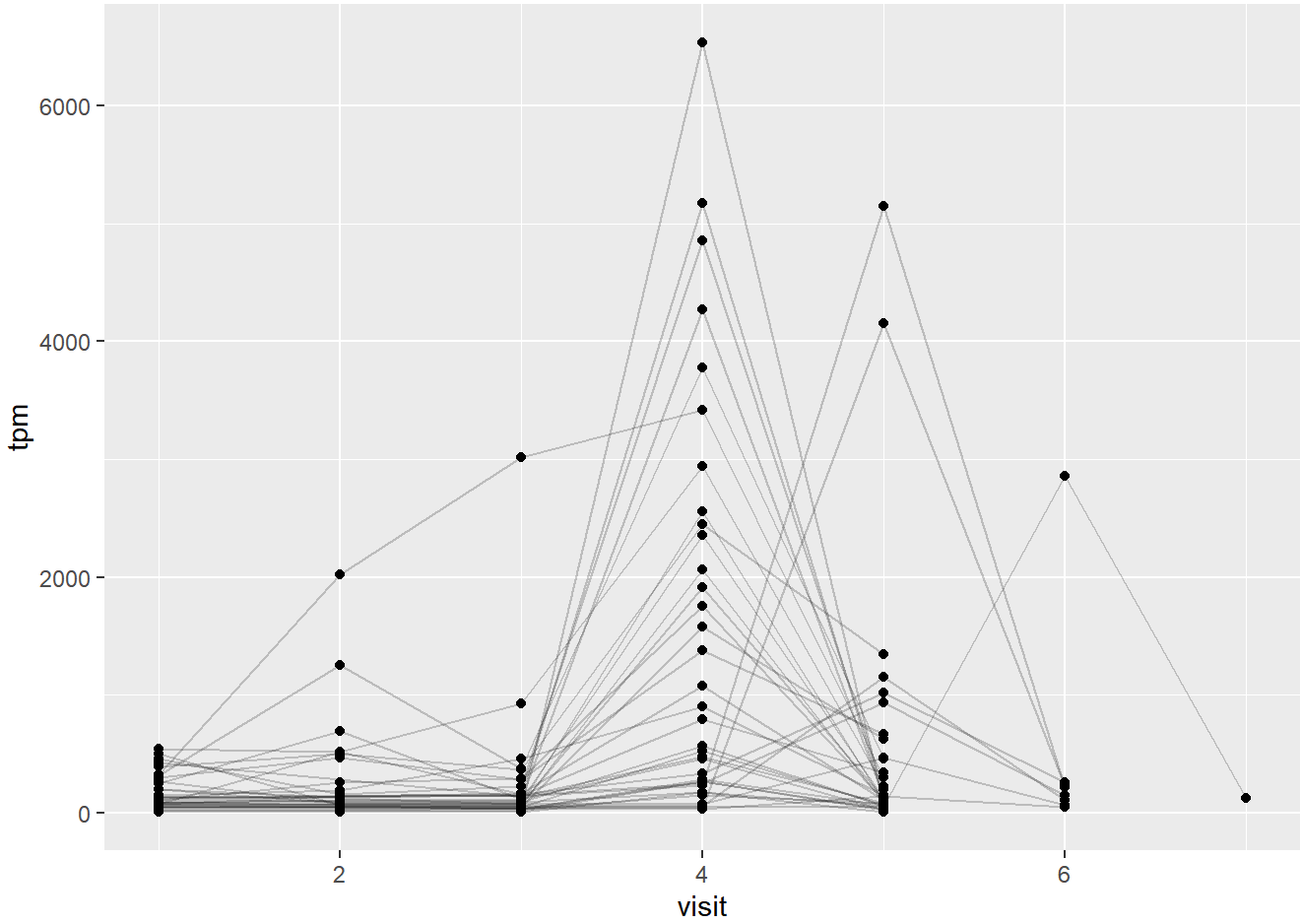
```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.7"

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).

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



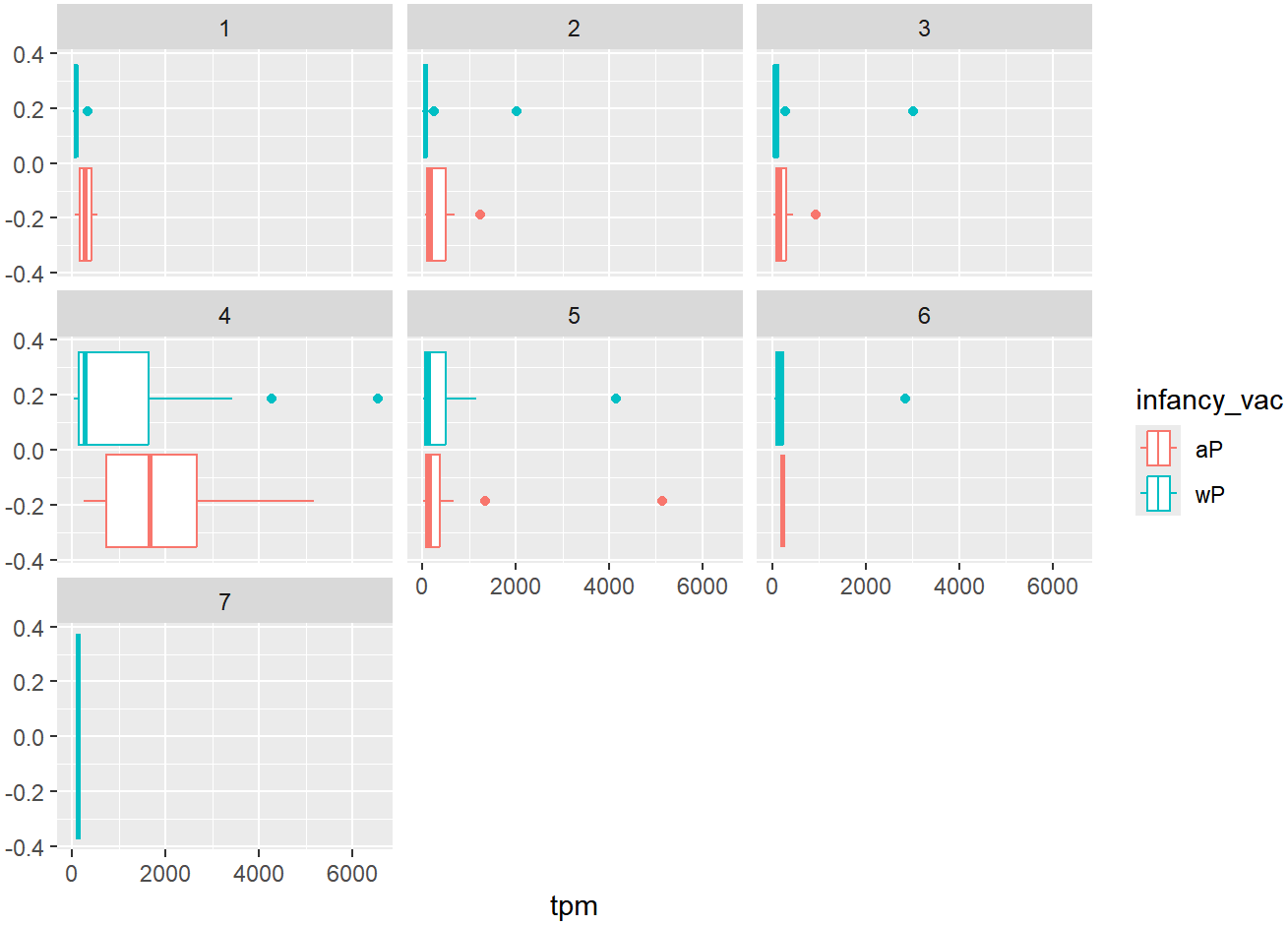
#Q20.: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

#The expression of this gene is at its maximum level at visit 4.

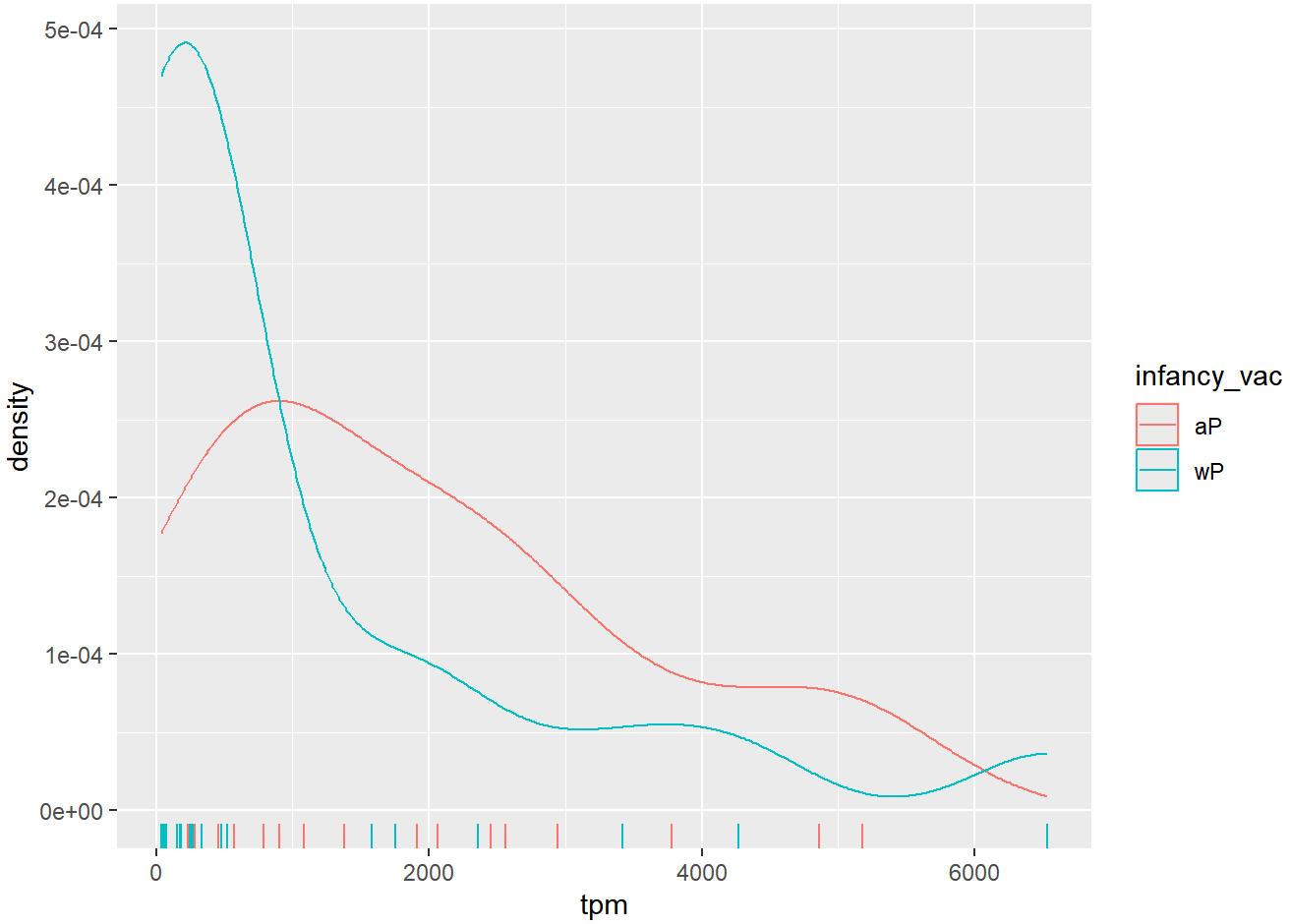
#Q21. Does this pattern in time match the trend of antibody titer data? If not, why not?

#No it does not match because the expression of the genes peak on different visit days.

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




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


5e-04
4e-04
3e-04
2e-04
1e-04
0e+00
density
0
2000
4000
6000
tpm
infancy_vac
aP
wP