Class 18: Pertussis Mini Project

Aileen Andrade (PID A17033749)

2025-03-09

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

The CDC tracks Pertussis cases around the U.S. https://tinyurl.com/pertussiscdc

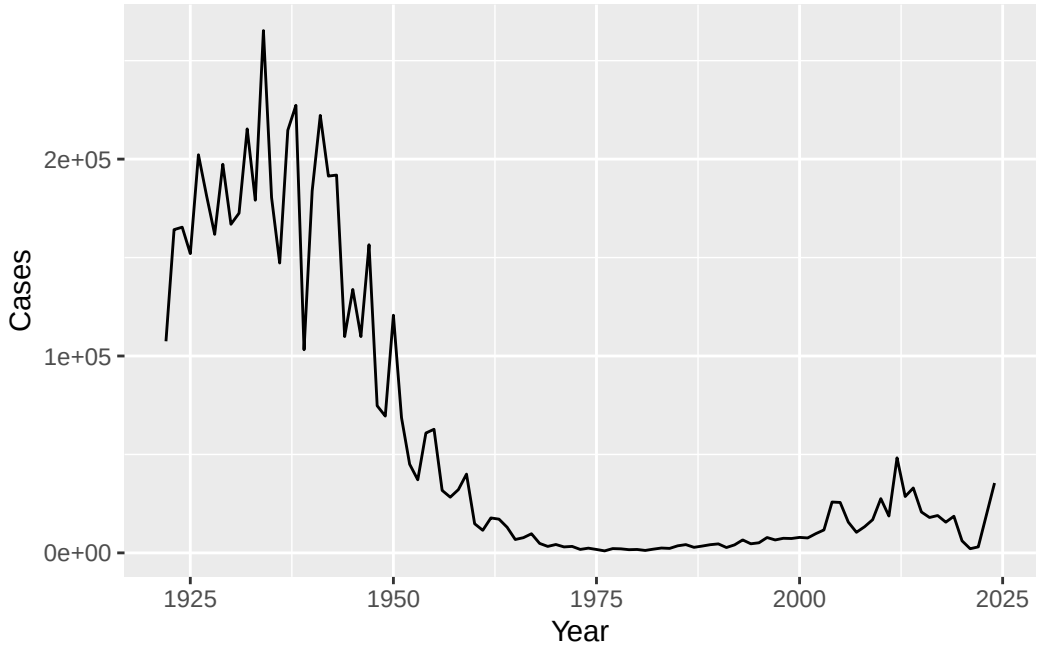
We can "scrape" this data using the R **datapasta** package.

```
head(cdc)
```

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

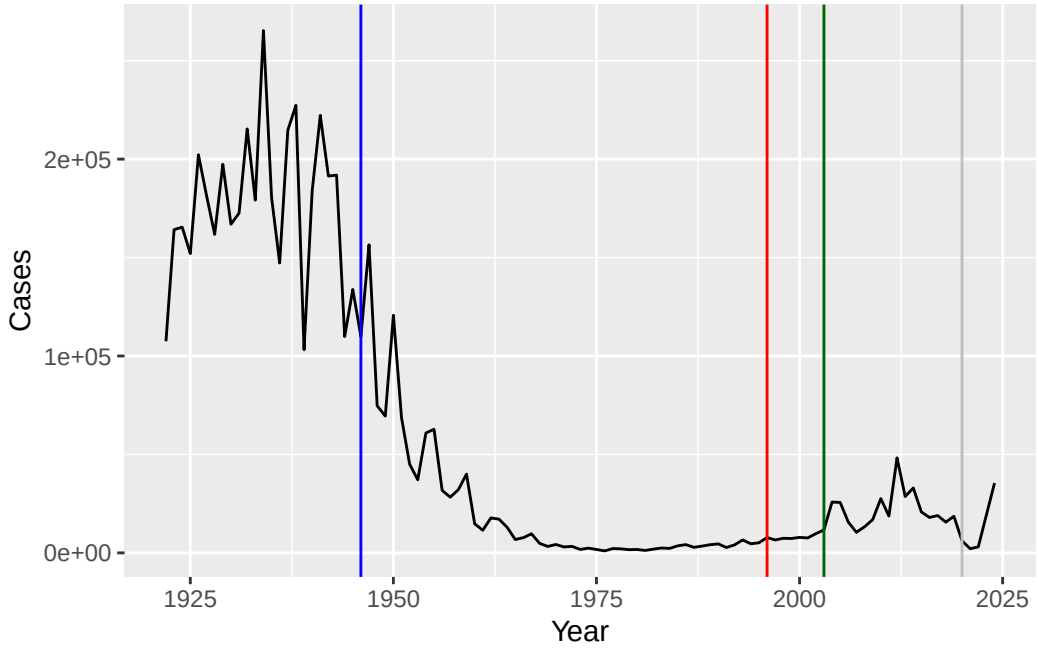
> Q1. With the help of the R "addin" package datapasta assign the CDC pertussis case number data to a data frame called cdc and use ggplot to make a plot of cases numbers over time.

```
library(ggplot2)

ggplot(cdc) +
  aes(Year, Cases) +
  geom_line()
```



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?

```
library(ggplot2)

ggplot(cdc) +
  aes(Year, Cases) +
  geom_line() +
  geom_vline(xintercept = 1946, col="Blue") +
  geom_vline(xintercept = 1996, col="Red") +
  geom_vline(xintercept = 2020, col="Gray") +
  geom_vline(xintercept = 2003, col="Dark green")
```



Observations: 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 outbreaks of pertussis again. There is also a notable COVID related dip and recent rapid rise.

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

After the introduction of the aP vaccine, in 1996 pertussis cases remained low initially until there was a resurgence starting around the 2000s. In 2012, there were 48,277 reported cases which was the highest since 1955. A possible explanation for the observed trend involves more sensitive PCR-based testing since the improved accuracy and sensitivity of the tests could have led to more detection as well as a decline in vaccines due to concerns. Additionally, bacterial evolution could have been a factor as well as the fact that the aP vaccine doesn't provide long-lasting immunity.

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

Computational Models of Immunity - Pertussis Boost CMI-PB

The CMI-PB project aims to address this key question: what is different 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")`

```
library(jsonlite)

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

Q. How many individuals "subjects" are in this data set?

```
nrow(subject)
```

```
[1] 172
```

Q4. How many wP and aP primed individuals are in this 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
```

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
```

```
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
```

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

subject$age_years <- time_length(subject$age, "years")

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
```

```
# Check if the difference is statistically significant using a t-test
t.test(ap$age_years, wp$age_years)
```

```
    Welch Two Sample t-test

data:  ap$age_years and wp$age_years
t = -12.918, df = 104.03, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.094058  -7.407351
sample estimates:
mean of x mean of y
 27.08358  35.83428
```

The average age of wP individuals is 35.83 years. The average age of aP individuals is 27.08 years. The difference is statistically significant (p-value $<$ 2.2e-16), meaning the wP group is significantly older than the aP group.

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

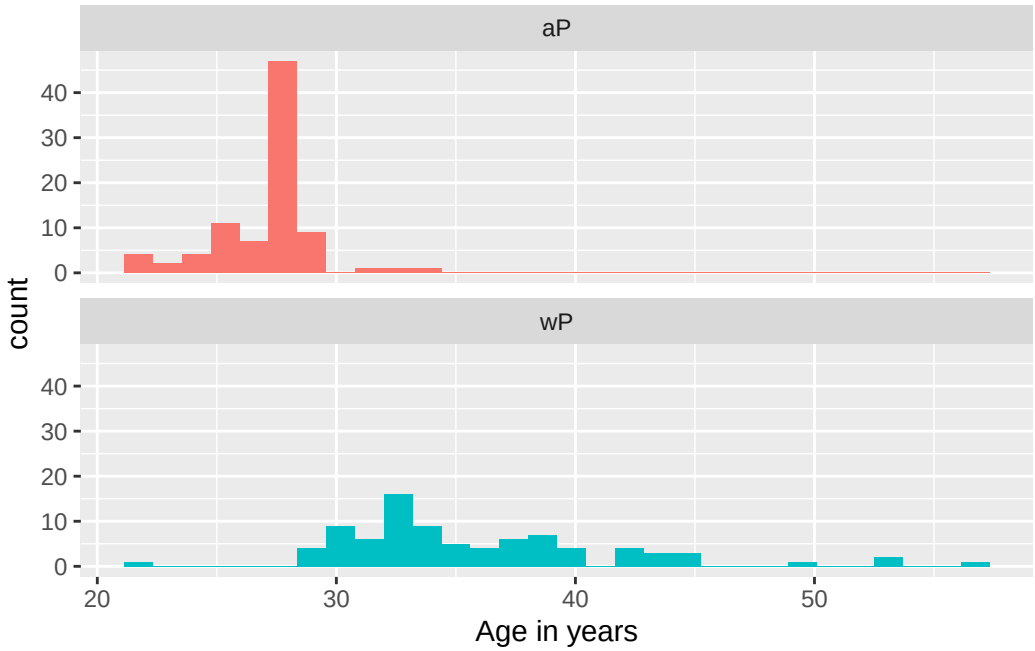
```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

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

Q9 (1). 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`.
```



With the help of a faceted histogram yes, the two groups are significantly different in age distrbution.

Obtain more data from CMI-PB:

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

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

```
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
```

> Q9 (2). 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:

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 **dplyr** package.

```
library(dplyr)

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

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

```
head(meta)
```

```
  subject_id infancy_vac biological_sex              ethnicity  race
1          1          wP         Female Not Hispanic or Latino White
2          1          wP         Female Not Hispanic or Latino White
3          1          wP         Female Not Hispanic or Latino White
4          1          wP         Female Not Hispanic or Latino White
5          1          wP         Female Not Hispanic or Latino White
6          1          wP         Female Not Hispanic or Latino White
  year_of_birth date_of_boost      dataset        age age_years specimen_id
1    1986-01-01    2016-09-12 2020_dataset 14313 days  39.18686           1
2    1986-01-01    2016-09-12 2020_dataset 14313 days  39.18686           2
3    1986-01-01    2016-09-12 2020_dataset 14313 days  39.18686           3
4    1986-01-01    2016-09-12 2020_dataset 14313 days  39.18686           4
5    1986-01-01    2016-09-12 2020_dataset 14313 days  39.18686           5
6    1986-01-01    2016-09-12 2020_dataset 14313 days  39.18686           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
```

```
dim(subject)
```

```
[1] 172  10
```

```
dim(specimen)
```

```
[1] 1503    6
```

```
dim(meta)
```

```
[1] 1503   15
```

> 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.

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

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

```
head(abdata)
```

```
  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 age_years specimen_id
1    1986-01-01    2016-09-12 2020_dataset 14313 days  39.18686           1
2    1986-01-01    2016-09-12 2020_dataset 14313 days  39.18686           1
3    1986-01-01    2016-09-12 2020_dataset 14313 days  39.18686           1
4    1986-01-01    2016-09-12 2020_dataset 14313 days  39.18686           1
5    1986-01-01    2016-09-12 2020_dataset 14313 days  39.18686           1
6    1986-01-01    2016-09-12 2020_dataset 14313 days  39.18686           1
  actual_day_relative_to_boost planned_day_relative_to_boost specimen_type
1                           -3                             0         Blood
2                           -3                             0         Blood
3                           -3                             0         Blood
4                           -3                             0         Blood
5                           -3                             0         Blood
6                           -3                             0         Blood
  visit isotype is_antigen_specific antigen        MFI MFI_normalised  unit
1     1     IgE               FALSE   Total 1110.21154       2.493425 UG/ML
2     1     IgE               FALSE   Total 2708.91616       2.493425 IU/ML
3     1     IgG                TRUE      PT   68.56614       3.736992 IU/ML
4     1     IgG                TRUE     PRN  332.12718       2.602350 IU/ML
5     1     IgG                TRUE     FHA 1887.12263      34.050956 IU/ML
6     1     IgE                TRUE     ACT    0.10000       1.000000 IU/ML
  lower_limit_of_detection
1                 2.096133
2                29.170000
3                 0.530000
4                 6.205949
5                 4.679535
6                 2.816431
```

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

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

```
[1] 61956
```

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  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?

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

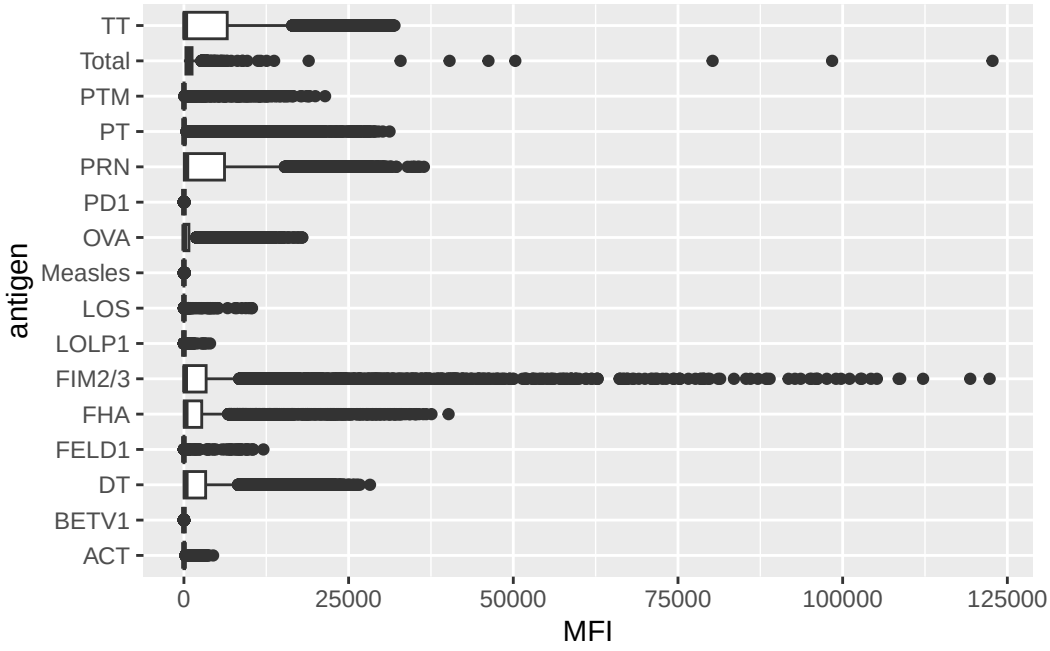
```
  ACT   BETV1      DT   FELD1     FHA  FIM2/3   LOLP1     LOS Measles     OVA
 1970    1970    6318    1970    6712    6318    1970    1970    1970    6318
  PD1     PRN      PT     PTM   Total      TT
 1970    6712    6712    1970     788    6318
```

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

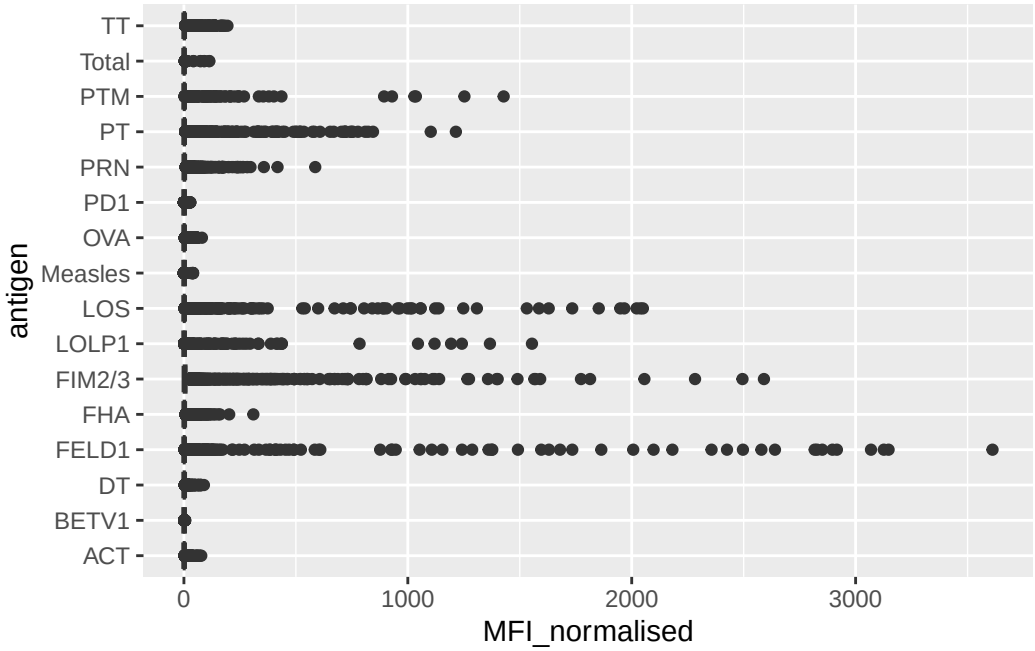
I want a plot of antigen levels across the whole dataset.

```
library(ggplot2)

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

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



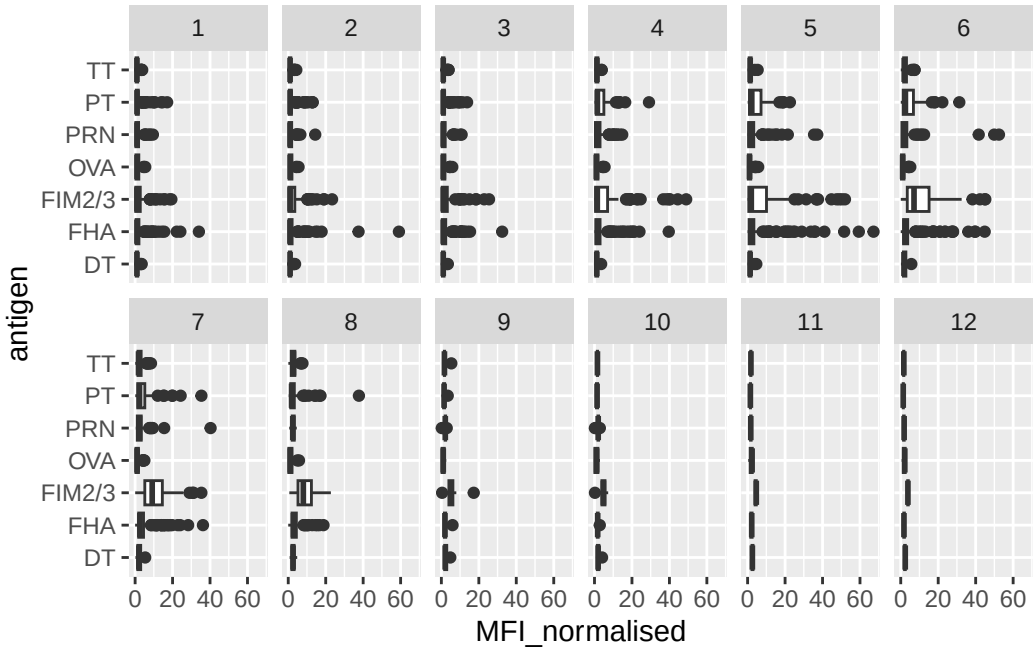
```
ggplot(abdata) +
  aes(MFI_normalised, antigen) +
  geom_boxplot()
```



Question 13 Summary Boxplot

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

```
igg_filtered <- igg %>% filter(MFI_normalised <= 75)

ggplot(igg_filtered) +
  aes(x = MFI_normalised, y = antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow = 2)
```

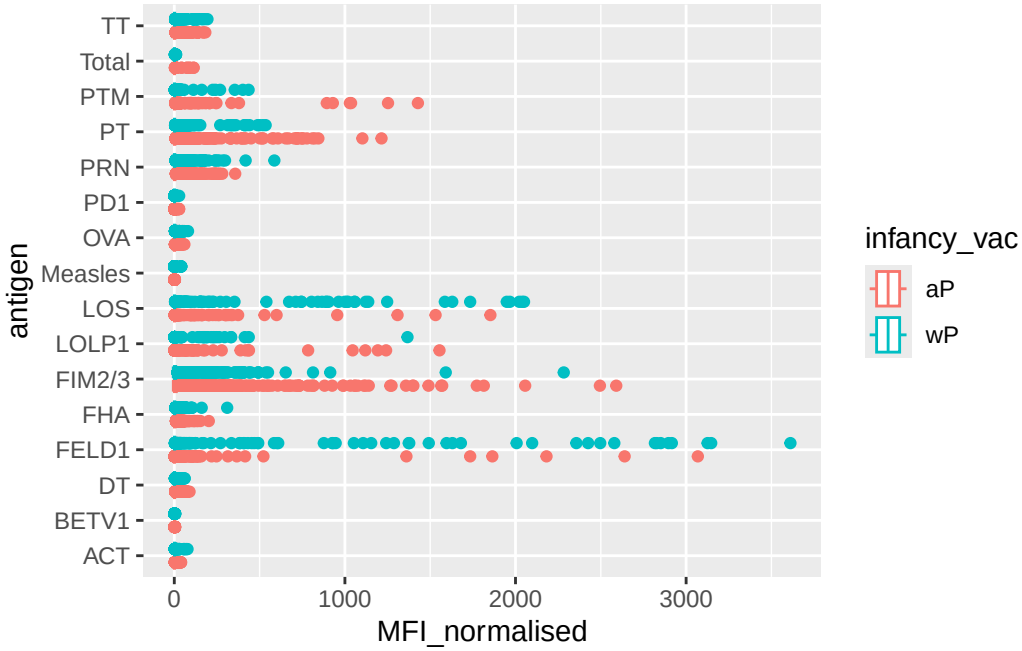


Q14. What antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others?

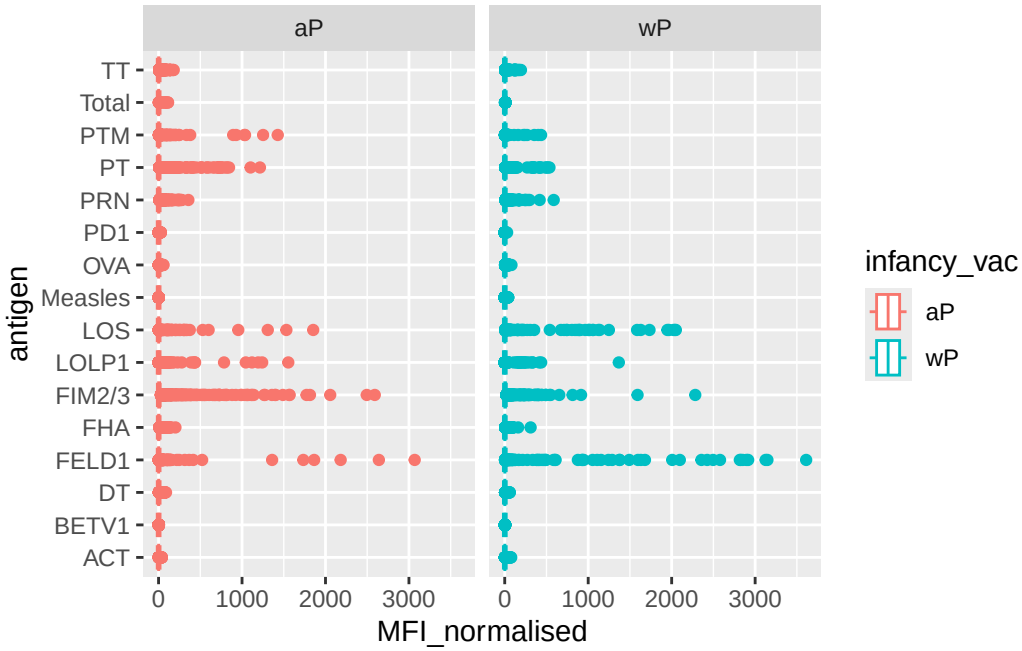
Antigens like FIM2/3, PT, FELD1 have quite a large range of values. Others like Measles don't show much activity. FIM2/3 and PT show differences since they're major components of the pertussis vaccines so individuals exposed to those vaccines can develop immune responses to them. Additionally, FELD1 is an antigen that has a strong immune reaction in certain individuals possibly due to previous infections or the environment. PT tends to also generate strong antibody responses in individuals with different vaccine types; measles doesn't show much activity since its antibody titers remain relatively stable as measles can induce long-lasting immunity from infection or vaccination.

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)
```



Examine IgG Ab titer levels

For this I need to select out just isotype IgG.

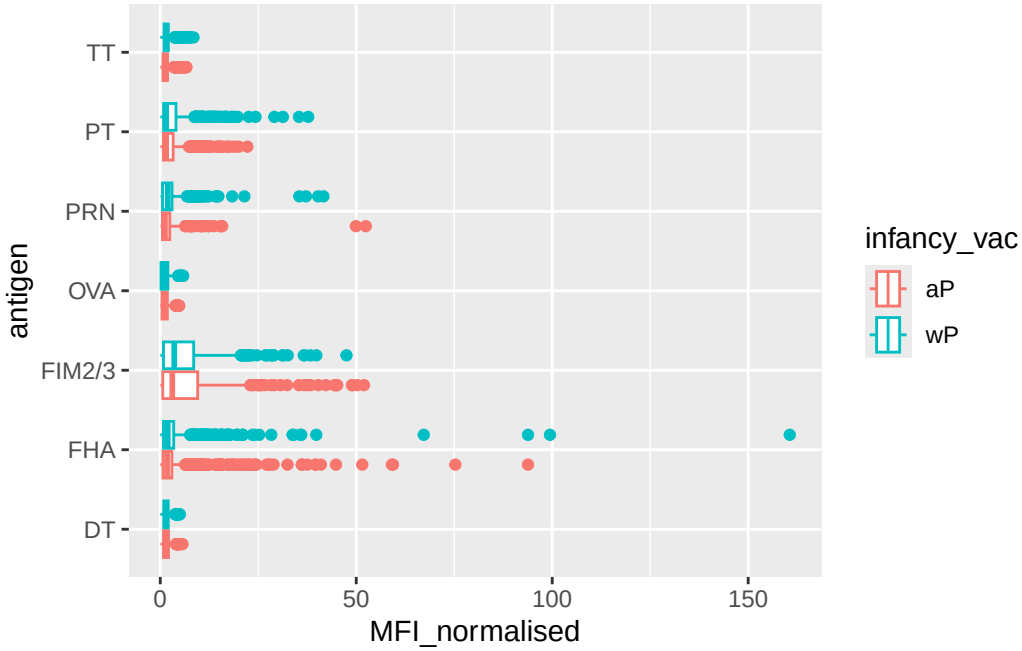
```
igg <- abdata |>
  filter(isotype == "IgG")
head(igg)
```

```
  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 age_years specimen_id
1    1986-01-01    2016-09-12 2020_dataset 14313 days  39.18686           1
2    1986-01-01    2016-09-12 2020_dataset 14313 days  39.18686           1
3    1986-01-01    2016-09-12 2020_dataset 14313 days  39.18686           1
4    1986-01-01    2016-09-12 2020_dataset 14313 days  39.18686           2
5    1986-01-01    2016-09-12 2020_dataset 14313 days  39.18686           2
6    1986-01-01    2016-09-12 2020_dataset 14313 days  39.18686           2
```

```
  actual_day_relative_to_boost planned_day_relative_to_boost specimen_type
1                           -3                             0         Blood
2                           -3                             0         Blood
3                           -3                             0         Blood
4                            1                             1         Blood
5                            1                             1         Blood
6                            1                             1         Blood
  visit isotype is_antigen_specific antigen        MFI MFI_normalised  unit
1     1     IgG                TRUE      PT   68.56614       3.736992 IU/ML
2     1     IgG                TRUE     PRN  332.12718       2.602350 IU/ML
3     1     IgG                TRUE     FHA 1887.12263      34.050956 IU/ML
4     2     IgG                TRUE      PT   41.38442       2.255534 IU/ML
5     2     IgG                TRUE     PRN  174.89761       1.370393 IU/ML
6     2     IgG                TRUE     FHA  246.00957       4.438960 IU/ML
  lower_limit_of_detection
1                 0.530000
2                 6.205949
3                 4.679535
4                 0.530000
5                 6.205949
6                 4.679535
```

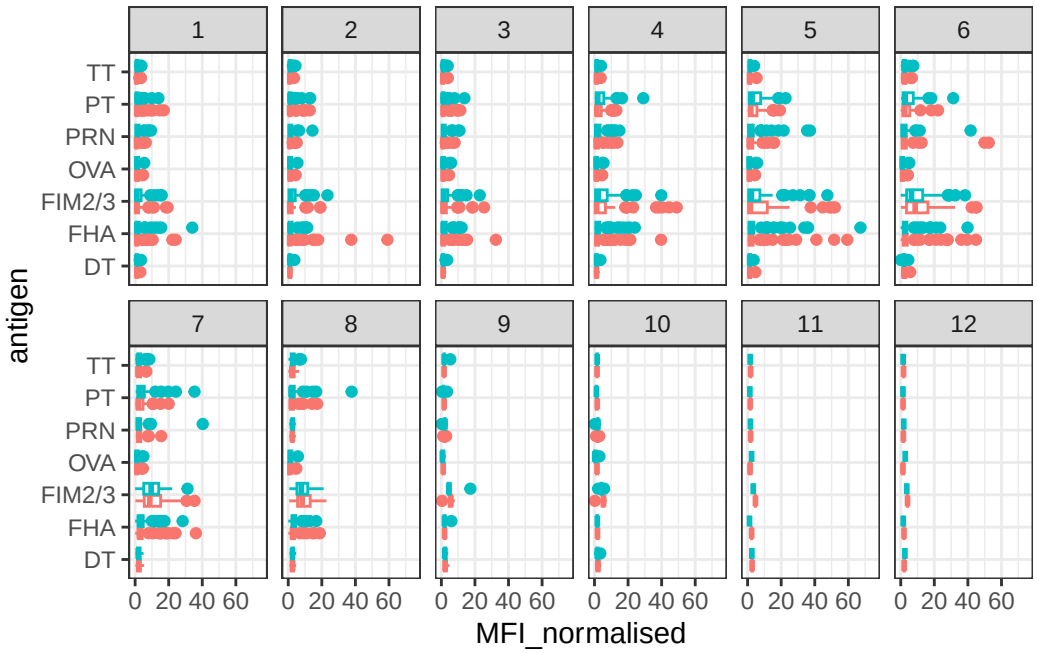
A overview boxplot:

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




```
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()
```

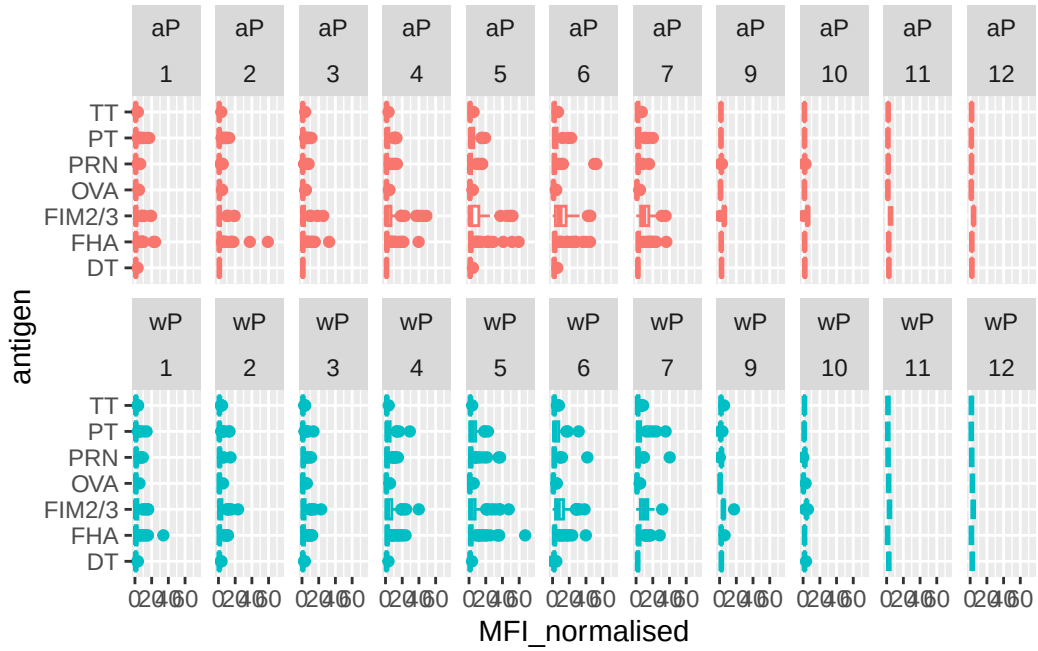
```
Warning: Removed 5 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



Another version of this plot adding infancy_vac to the faceting:

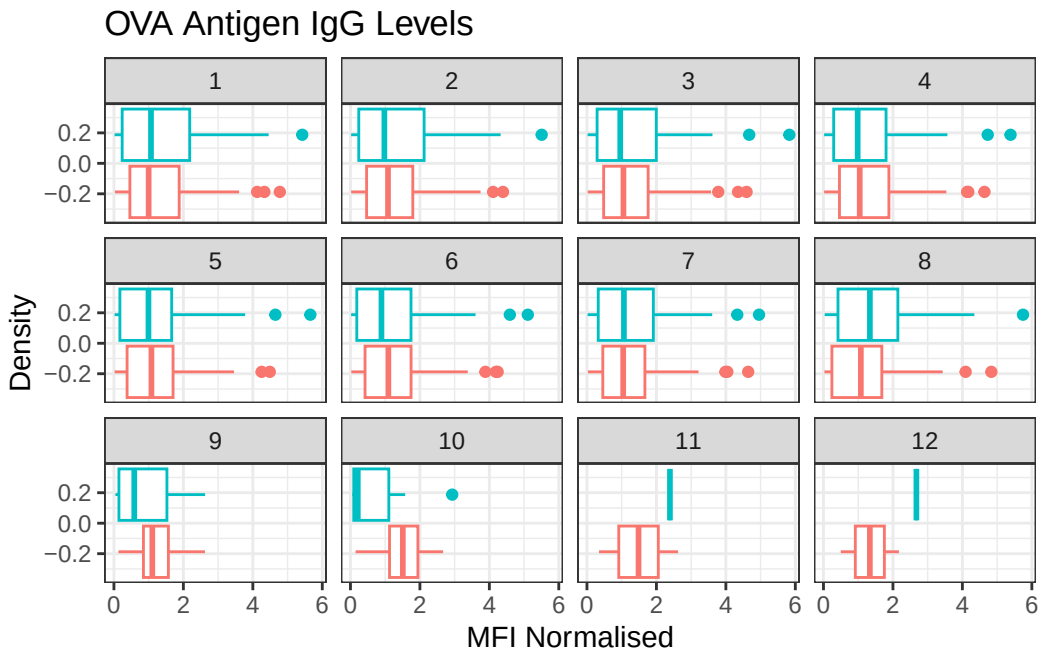
```
igg %>% filter(visit != 8) %>%
ggplot() +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  xlim(0,75) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```

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

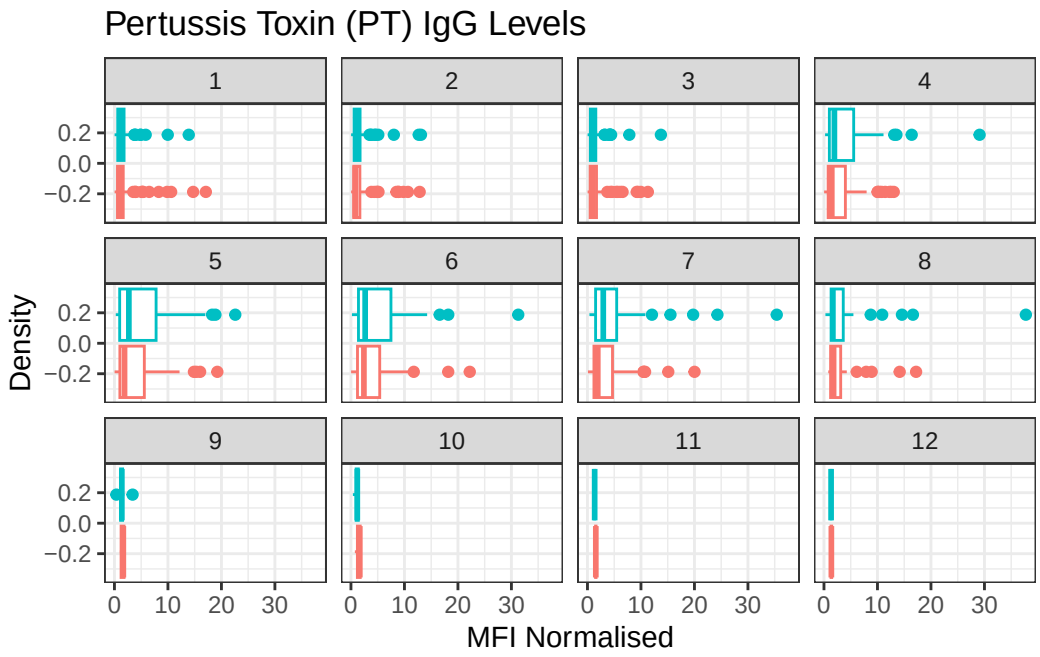


Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("OVA", that is not in our vaccines) and a clear antigen of interest ("PT", Pertussis Toxin, one of the key virulence factors produced by the bacterium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

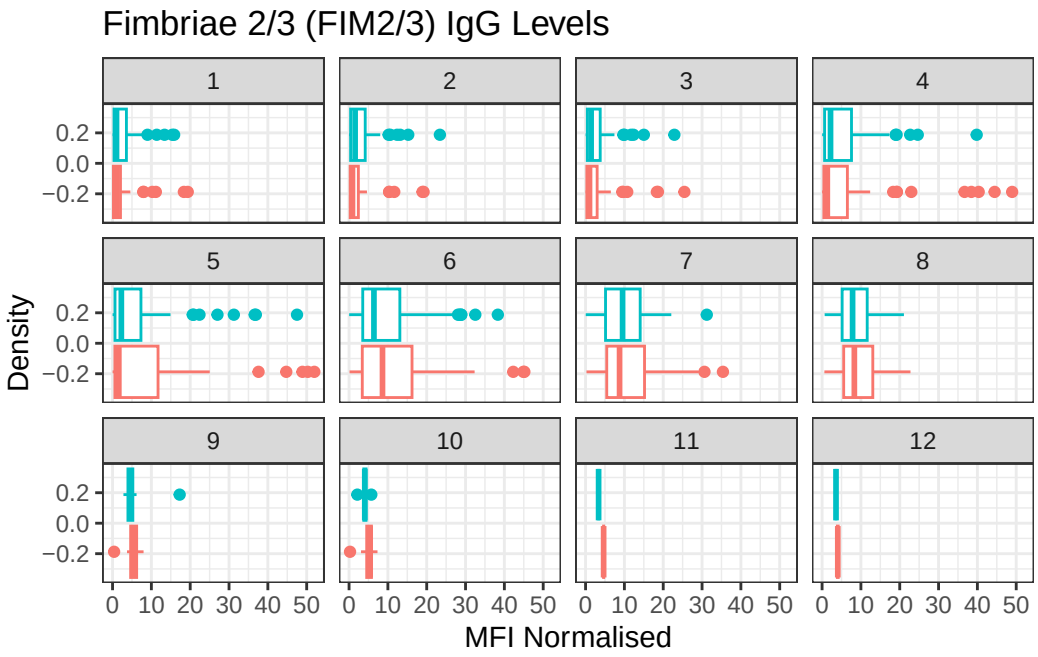
```
library(dplyr)
library(ggplot2)

# Boxplot for OVA (Control Antigen - not in vaccine)
filter(igg, antigen == "OVA") %>%
  ggplot() +
  aes(x = MFI_normalised, col = infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw() +
  labs(title = "OVA Antigen IgG Levels", x = "MFI Normalised", y = "Density")
```




```
# Boxplot for PT (Pertussis Toxin - key virulence factor)
filter(igg, antigen == "PT") %>%
  ggplot() +
  aes(x = MFI_normalised, col = infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw() +
  labs(title = "Pertussis Toxin (PT) IgG Levels", x = "MFI Normalised", y = "Density")
```




```
# Boxplot for FIM2/3 (Fimbriae 2/3 - Pertussis Vaccine Component)
filter(igg, antigen == "FIM2/3") %>%
  ggplot() +
  aes(x = MFI_normalised, col = infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw() +
  labs(title = "Fimbriae 2/3 (FIM2/3) IgG Levels", x = "MFI Normalised", y = "Density")
```



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

I notice that OVA titers remain low and stable which confirms its a control antigen; it is also not present in the pertussis vaccine so theres no immune response to it. PT titers rise, peak, and decline showing an immune response to a vaccine antigen. Lastly, the wP and aP subjects pattern is similar, both vaccine types induce a response to PT but the differences can be analyzed further.

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

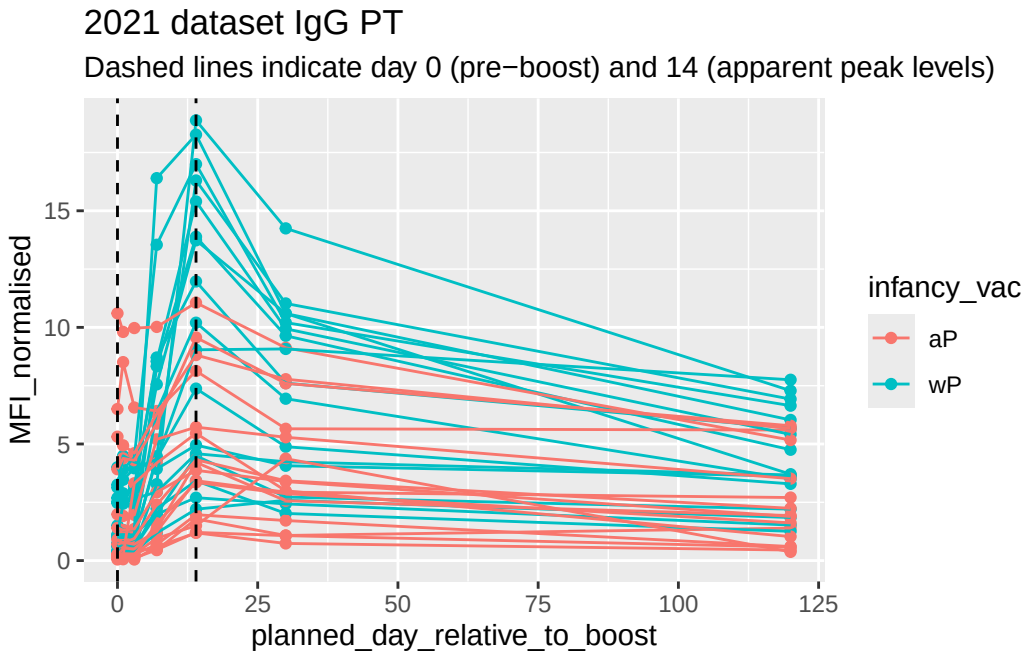
There are clear differences, aP individuals have a higher initial IgG titer response but the levels decline faster. wP individuals have more sustained IgG levels with potentially longer-lasting immune responses than aP individuals.

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

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

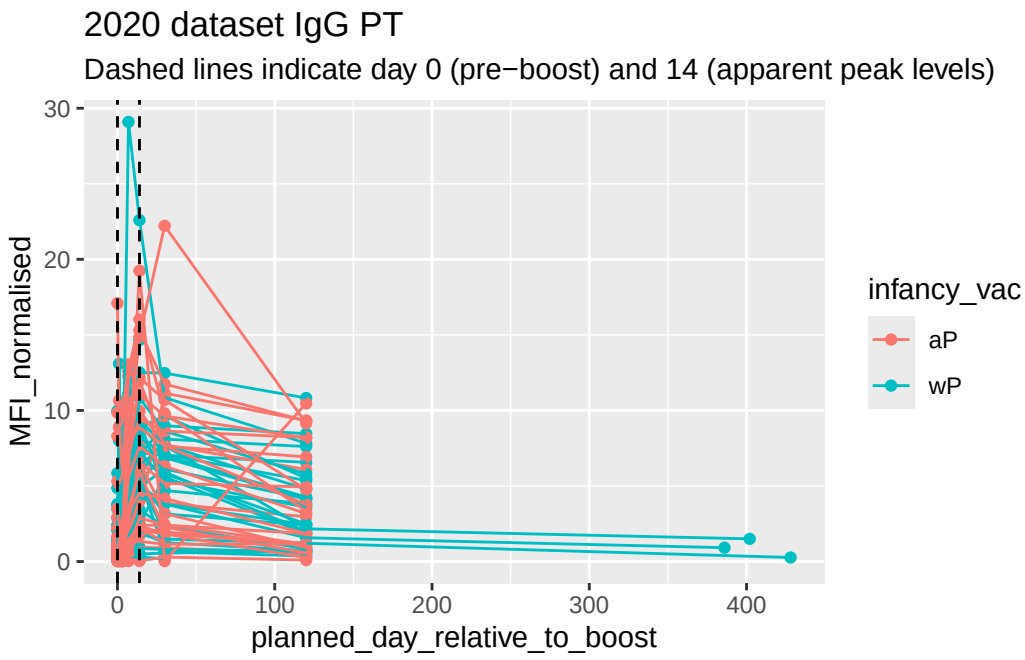
```
## Filter to include 2021 data only
abdata.21 <- abdata |>
  filter(dataset == "2021_dataset")
```

```r
## 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)")
```




```r
# Filter for the 2020 dataset
abdata.20 <- abdata %>% filter(dataset == "2020_dataset")

# Plot time-course for IgG PT antigen levels in 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)")
```



Yes the overall trend is similar between the 2020 and 2021 datasets. Both show peaks at day 14 post-boost followed by a decline. What is most notable is that the 2020 dataset extends for a longer duration which displays antibody levels continuing to decrease over time. Across both years, the higher response in wP individuals compared to aP individuals is also consistent.

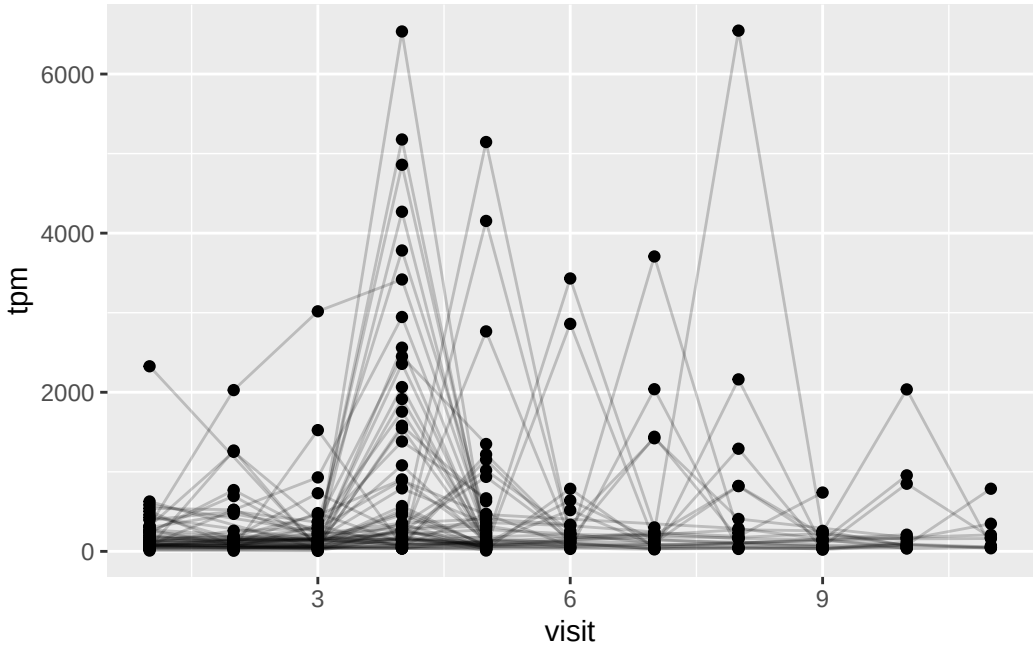
Q19. Make a plot of the time course of gene expression for IGHG1 gene (i.e. a plot of visit vs. tpm).

```
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)`
```

```
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)?

From the plot, IGHG1 expression (TPM values) peak around visit 4 and visit 8. This can indicate two waves of B cell activation or expansion; the first peak around visit 4 can represent the initial activation of B cells and early surge in IgG1 transcription whereas the second peak around visit 8 can be secondary activation, memory B cell responses.

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

No, the pattern of IGHG1 expression over time does not exactly match the trend of antibody titer data from question 15. This is because IGHG1 gene expression peaks early around visit 4 and visit 8 and it declines quickly after the peaks. Antibody titer levels in question 15 show IgG levels rise post-boost peaking around visit 4-6. They also remain high for a long period evven when IGHG1 transcription decreases. There might be a mismatch because IGHG1 expression is needed early and B cells need to transcribe the IGHG1 gene before it can start producing IgG1 antibodies. The high levels of IgG antibody titers in question 15 represent plasma cells as they continue producing antibodies for a long time, even after IGHG1 mRNA levels drop. So while IgG antibody levels remain elevated for a longer time, IGHG1 expression peaks early and decline since plasma cells continue producing IgG antibodies without needing high IGHG1 expression. Even after IGHG1 expression decreases, titer levels stay high since antibodies persist in the bloodstream for weeks to months.