18._Pertussis_Resurgence_(mini-project)

Edwin Ruiz

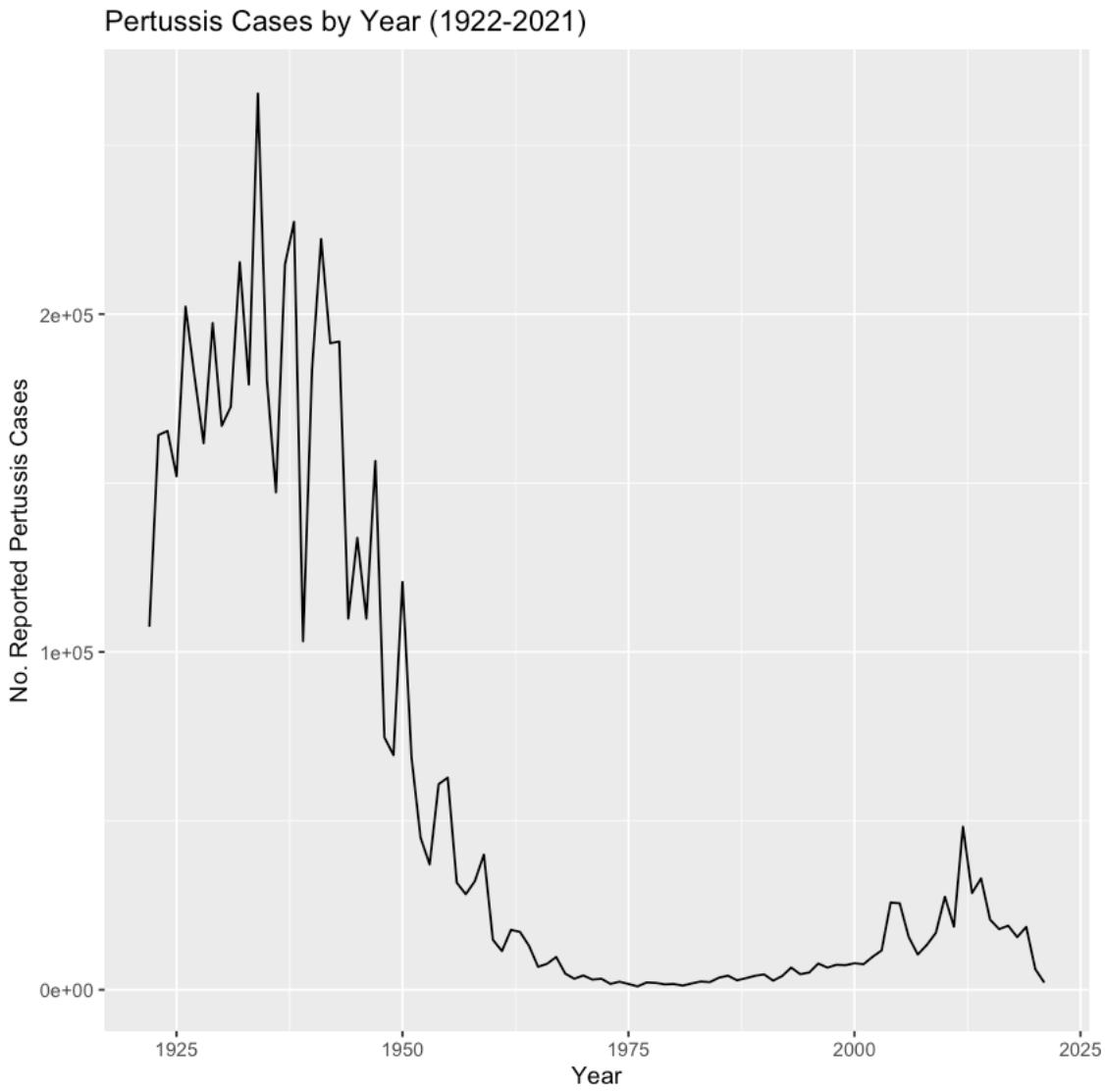
1 Pertussis and the CMI-PB project

1.1 1. Investigating pertussis cases by year

[1]:
```
library(ggplot2)
cdc <- read.csv("Pertussis_Cases_by_Year_1922-2021.csv")

names(cdc) <- c("Year", "Cases")

ggplot(cdc, aes(x = Year, y = Cases)) +
  geom_line() +
  labs(title = "Pertussis Cases by Year (1922-2021)",
       x = "Year",
       y = "No. Reported Pertussis Cases")
```



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

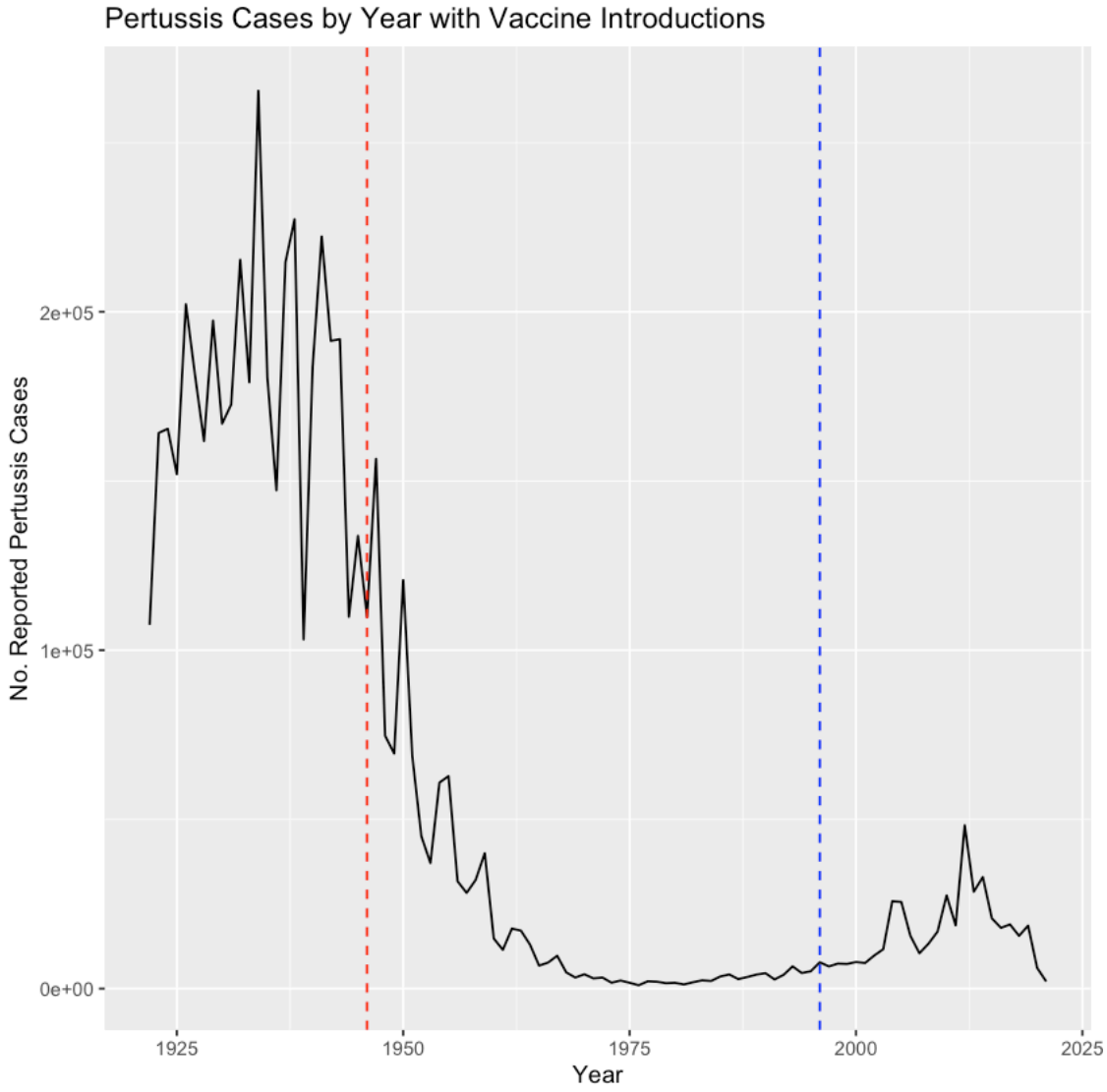
Q2. Using the ggplot geom_vline() function add lines to your previous plot for the 1946 introduction of the wP vaccine and the 1996 switch to aP vaccine (see example in the hint below). What do you notice?

1.3 I notice an increase in the incidence of pertussis cases from the early 2000s onward.

[3]:
```
library(ggplot2)

ggplot(cdc, aes(x = Year, y = Cases)) +
  geom_line() +
```

```
  geom_vline(xintercept = 1946, linetype = "dashed", color = "red") +
  geom_vline(xintercept = 1996, linetype = "dashed", color = "blue") +
  labs(title = "Pertussis Cases by Year with Vaccine Introductions",
       x = "Year",
       y = "No. Reported Pertussis Cases")
```



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

1.5 This new vaccine might not be as effective as the first.

2 3. Exploring CMI-PB data

[4]:
```
library(jsonlite)
subject <- read_json("https://www.cmi-pb.org/api/subject", simplifyVector =␣
 ↪TRUE)
head(subject, 3)
```

A data.frame: 3 × 8

	subject_id <int>	infancy_vac <chr>	biological_sex <chr>	ethnicity <chr>	race <chr>	year_of_ <chr>
1	1	wP	Female	Not Hispanic or Latino	White	1986-01-
2	2	wP	Female	Not Hispanic or Latino	White	1968-01-
3	3	wP	Female	Unknown	White	1983-01-

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

[5]:
```
ap_count <- sum(subject$infancy_vac == "aP")
print(paste("Number of subjects vaccinated with aP:", ap_count))
wp_count <- sum(subject$infancy_vac == "wP")
print(paste("Number of subjects vaccinated with wP:", wp_count))
```

```
[1] "Number of subjects vaccinated with aP: 60"
[1] "Number of subjects vaccinated with wP: 58"
```

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

[6]:
```
male_count <- sum(subject$biological_sex == "Male")
print(paste("Number of male subjects:", male_count))
female_count <- sum(subject$biological_sex == "Female")
print(paste("Number of female subjects:", female_count))
```

```
[1] "Number of male subjects: 39"
[1] "Number of female subjects: 79"
```

2.3 Q6. What is the breakdown of race and biological sex (e.g. number of Asian females, White males etc...)?

[7]:
```
table(subject$race, subject$biological_sex)
```

```
                                   Female Male
  American Indian/Alaska Native         0    1
  Asian                                21   11
  Black or African American             2    0
  More Than One Race                    9    2
```

```
  Native Hawaiian or Other Pacific Islander      1    1
  Unknown or Not Reported                       11    4
  White                                         35   20
```

[8]:
```
library(lubridate)
today()
```

```
Attaching package: ‘lubridate’


The following objects are masked from ‘package:base’:

    date, intersect, setdiff, union
```

2024-06-07

[9]:
```
today() - ymd("2000-01-01")
```

```
Time difference of 8924 days
```

[10]:
```
time_length( today() - ymd("2000-01-01"), "years")
```

24.4325804243669

3 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?

Yes they are significant.

[11]:
```
library(lubridate)
subject$year_of_birth <- ymd(subject$year_of_birth)
subject$age <- as.numeric(difftime(today(), subject$year_of_birth, units =␣
 ↪"days")) / 365.25

avg_age_wp <- mean(subject$age[subject$infancy_vac == "wP"], na.rm = TRUE)
avg_age_ap <- mean(subject$age[subject$infancy_vac == "aP"], na.rm = TRUE)
difference_significant <- t.test(subject$age ~ subject$infancy_vac)$p.value < 0.
 ↪05

t_test_result <- t.test(subject$age ~ subject$infancy_vac)
difference_significant <- t_test_result$p.value < 0.05
p_value <- t_test_result$p.value

print(paste("(i) the average age of wP individuals:", avg_age_wp))
```

```
print(paste("(ii) the average age of aP individuals:", avg_age_ap))
print(paste("(iii) are they significantly different? ",␣
 ↪ifelse(difference_significant, "Yes", "No")))
print(paste("P-value for (iii) significance:", format(p_value, scientific =␣
 ↪FALSE)))
```

```
[1] "(i) the average age of wP individuals: 36.8280582501357"
[1] "(ii) the average age of aP individuals: 26.5313255760894"
[1] "(iii) are they significantly different?  Yes"
[1] "P-value for (iii) significance: 0.000000000000000006813505"
```

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

[12]:
```
custom_round <- function(x) {
  floor(x + 0.4)
}

subject$date_of_boost <- ymd(subject$date_of_boost)
subject$age_at_boost <- as.numeric(difftime(subject$date_of_boost,␣
 ↪subject$year_of_birth, units = "days")) / 365.25
subject$age_at_boost <- sapply(subject$age_at_boost, custom_round)
print(head(subject$age_at_boost))
```

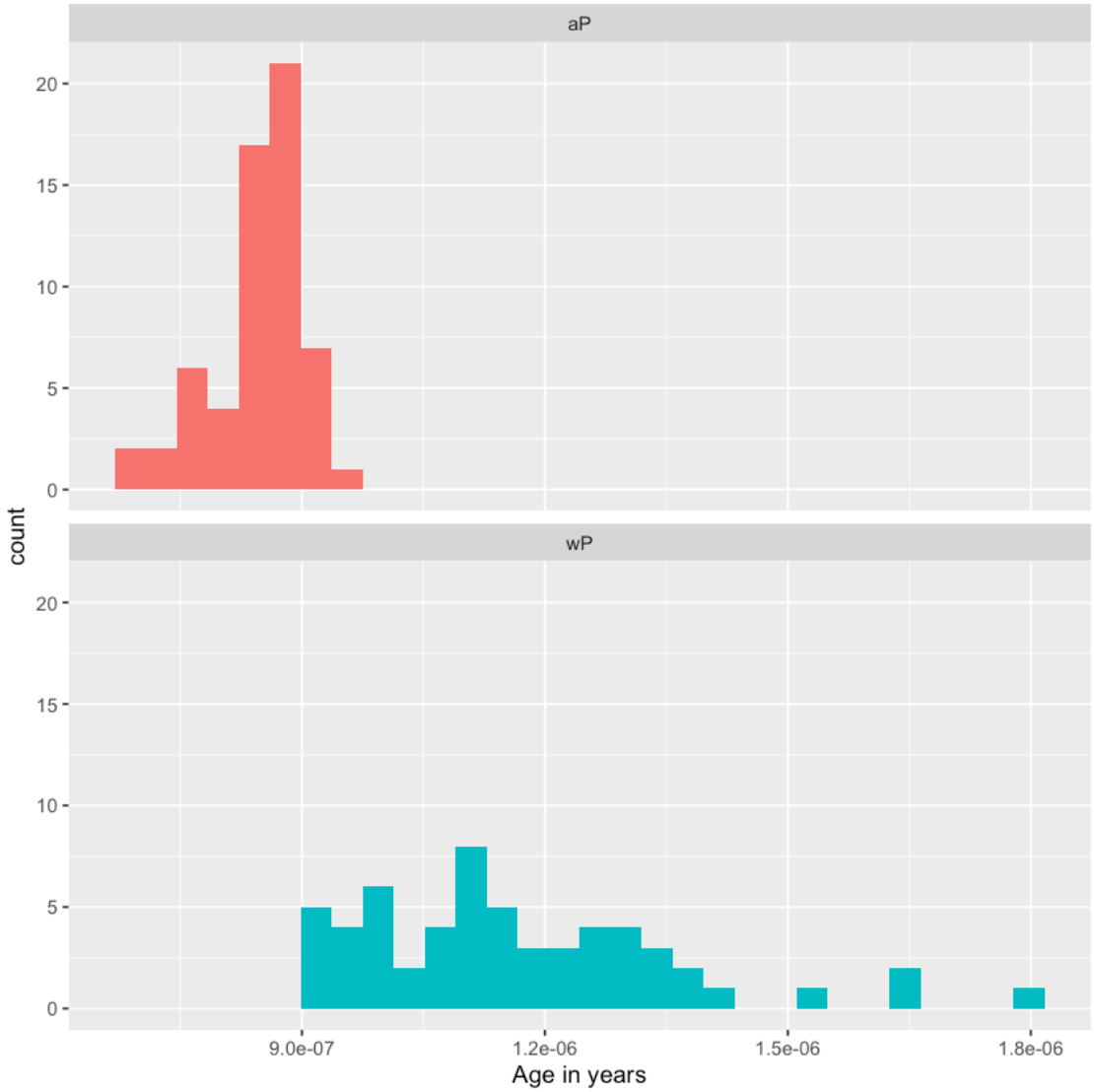
```
[1] 31 51 34 29 26 29
```

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

4 Yes they are significantly different, there is a separation in age distributions

[13]:
```
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`.
```



[15]:
```
specimen <- read_json("https://www.cmi-pb.org/api/specimen", simplifyVector =␣
↪TRUE)
titer <- read_json("https://www.cmi-pb.org/api/plasma_ab_titer", simplifyVector␣
↪= TRUE)
```

4.1 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:

[20]:
```
library(dplyr)
meta <- left_join(specimen, subject)
dim(meta)
```

```
head(meta)
```

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

1. 939 2. 13

A data.frame: 6 × 13

	specimen_id <int>	subject_id <int>	actual_day_relative_to_boost <int>	planned_day_relative_t <int>
1	1	1	-3	0
2	2	1	1	1
3	3	1	3	3
4	4	1	7	7
5	5	1	11	14
6	6	1	32	30

4.2 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.

[23]:
```
abdata <- inner_join(titer, meta)
dim(abdata)
```

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

1. 46906 2. 20

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

[26]:
```
table(abdata$isotype)
# isotype_counts <- table(abdata$isotype)
# isotype_counts
```

```
 IgE  IgG IgG1 IgG2 IgG3 IgG4
6698 4255 8983 8990 8990 8990
```

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

[27]:
```
dataset_counts <- table(abdata$dataset)
dataset_counts
```

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

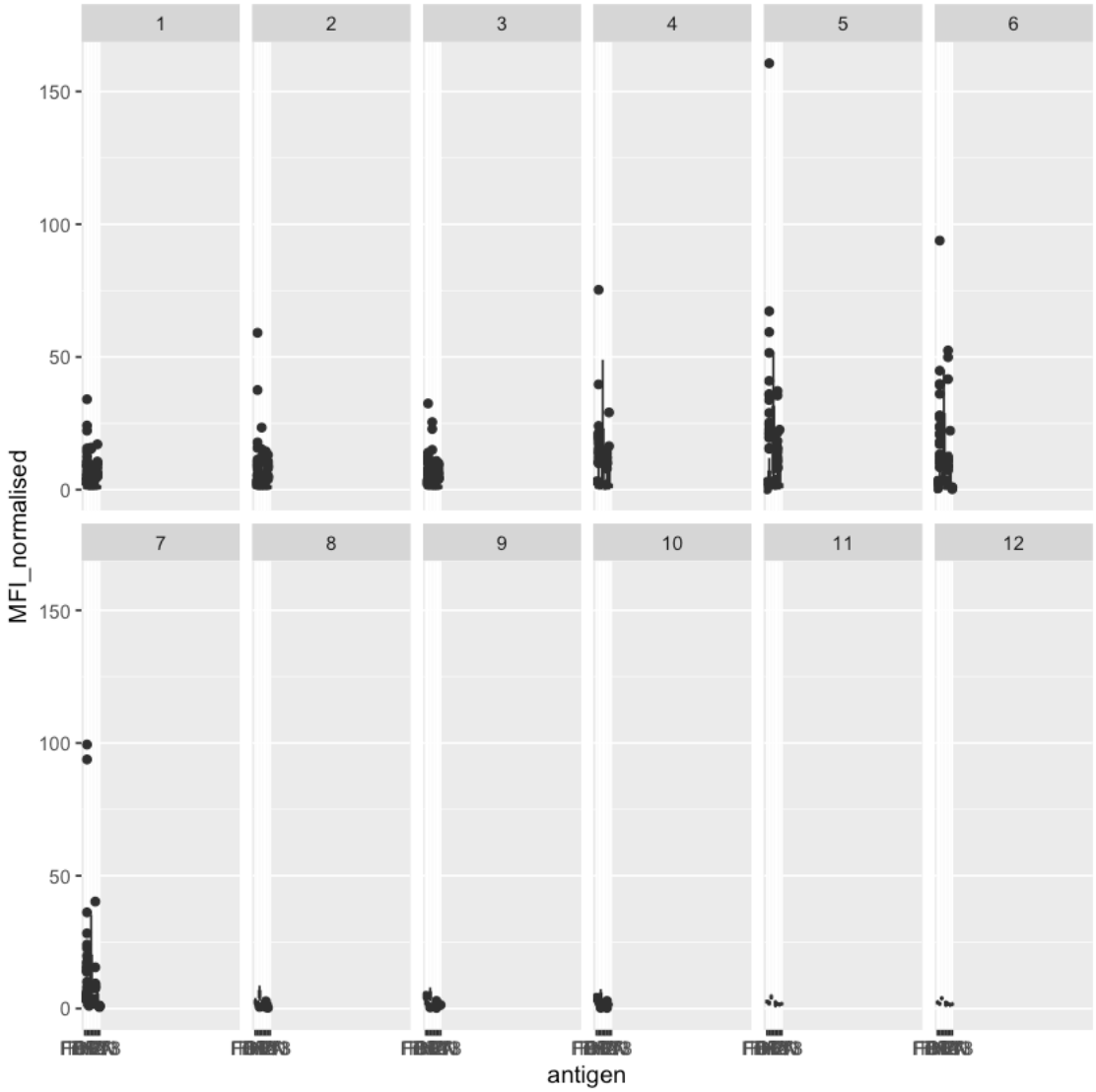
[28]:
```
igg <- abdata %>% filter(isotype == "IgG")
head(igg)
```

A data.frame: 6 × 20

	specimen_id <int>	isotype <chr>	is_antigen_specific <lgl>	antigen <chr>	MFI <dbl>	MFI_normalised <dbl>
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

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

[33]:
```
ggplot(igg) +
  aes(x=antigen, y=MFI_normalised) +
  geom_boxplot() +
  coord_cartesian(xlim=c(0, 75)) +
  facet_wrap(~visit, nrow=2)
```

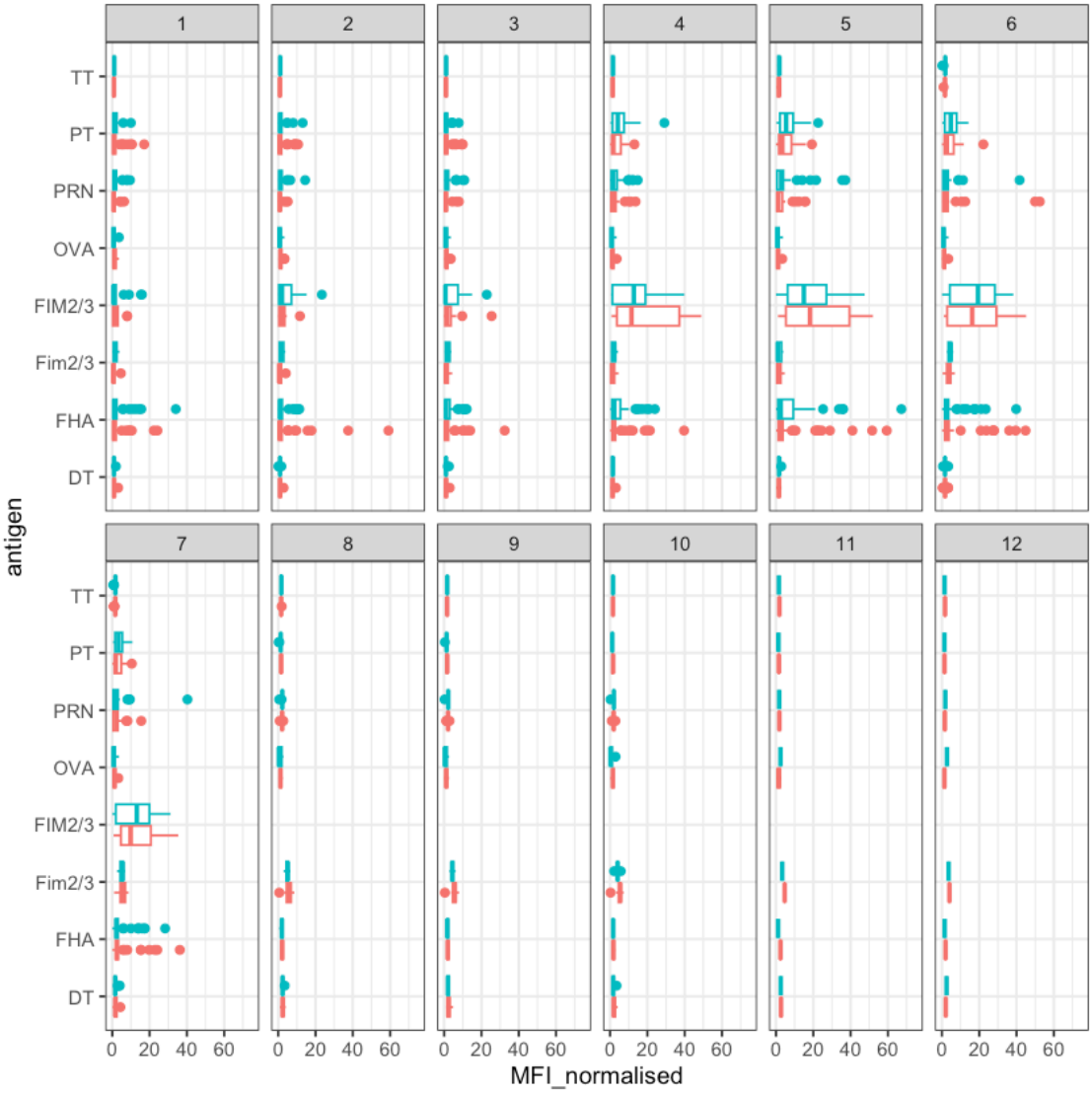


4.6 Q14. What antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others?

Antigens that show differences in IgG antibody titers over time are FIM2/3, FHA, PT, PRN, which show strong immune responses in that order due to the substances producing the immune responses.

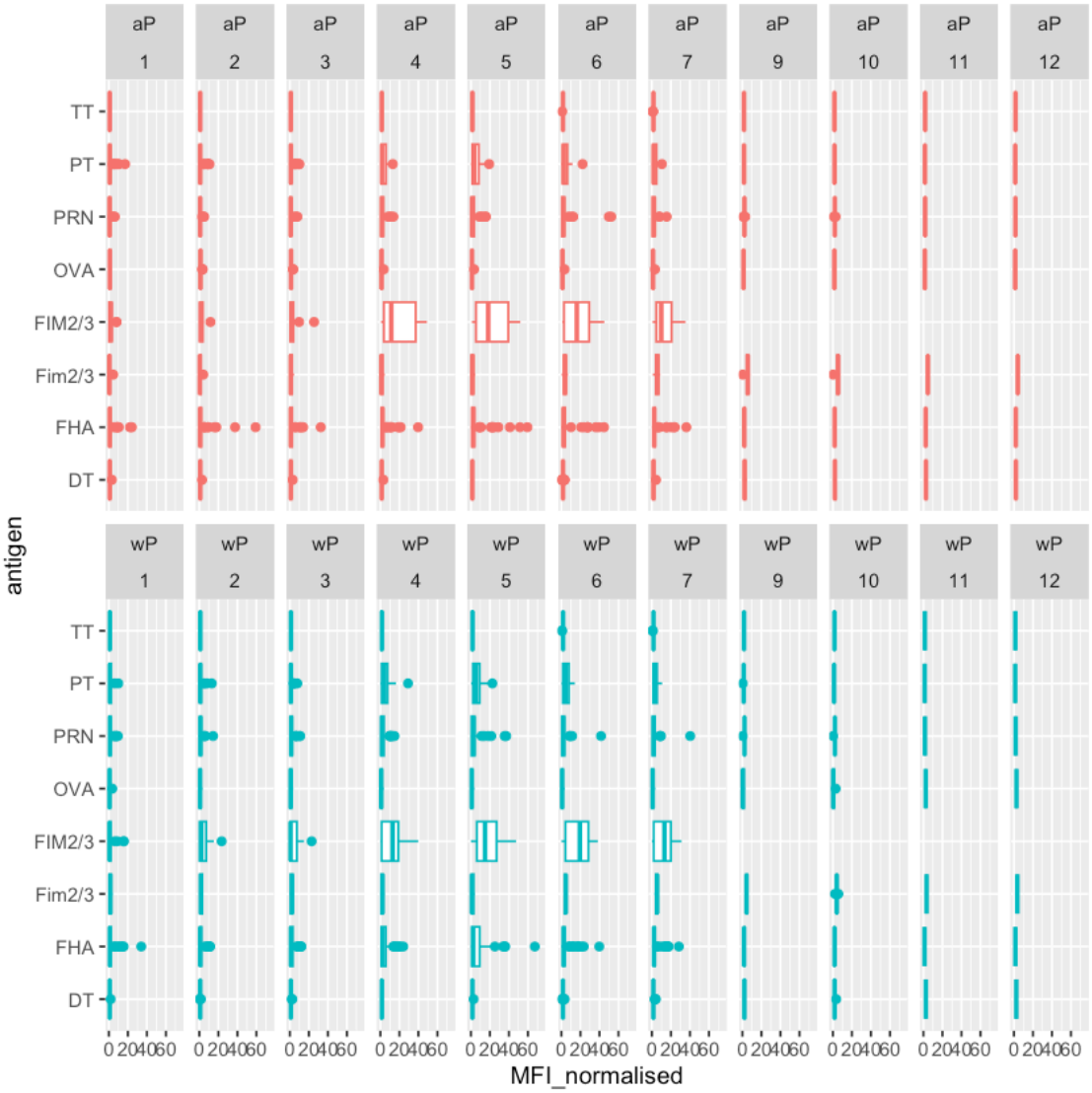
[34]:
```
ggplot(igg) +
aes(MFI_normalised, antigen, col=infancy_vac ) +
geom_boxplot(show.legend = FALSE) +
facet_wrap(vars(visit), nrow=2) +
xlim(0,75) +
theme_bw()
```

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



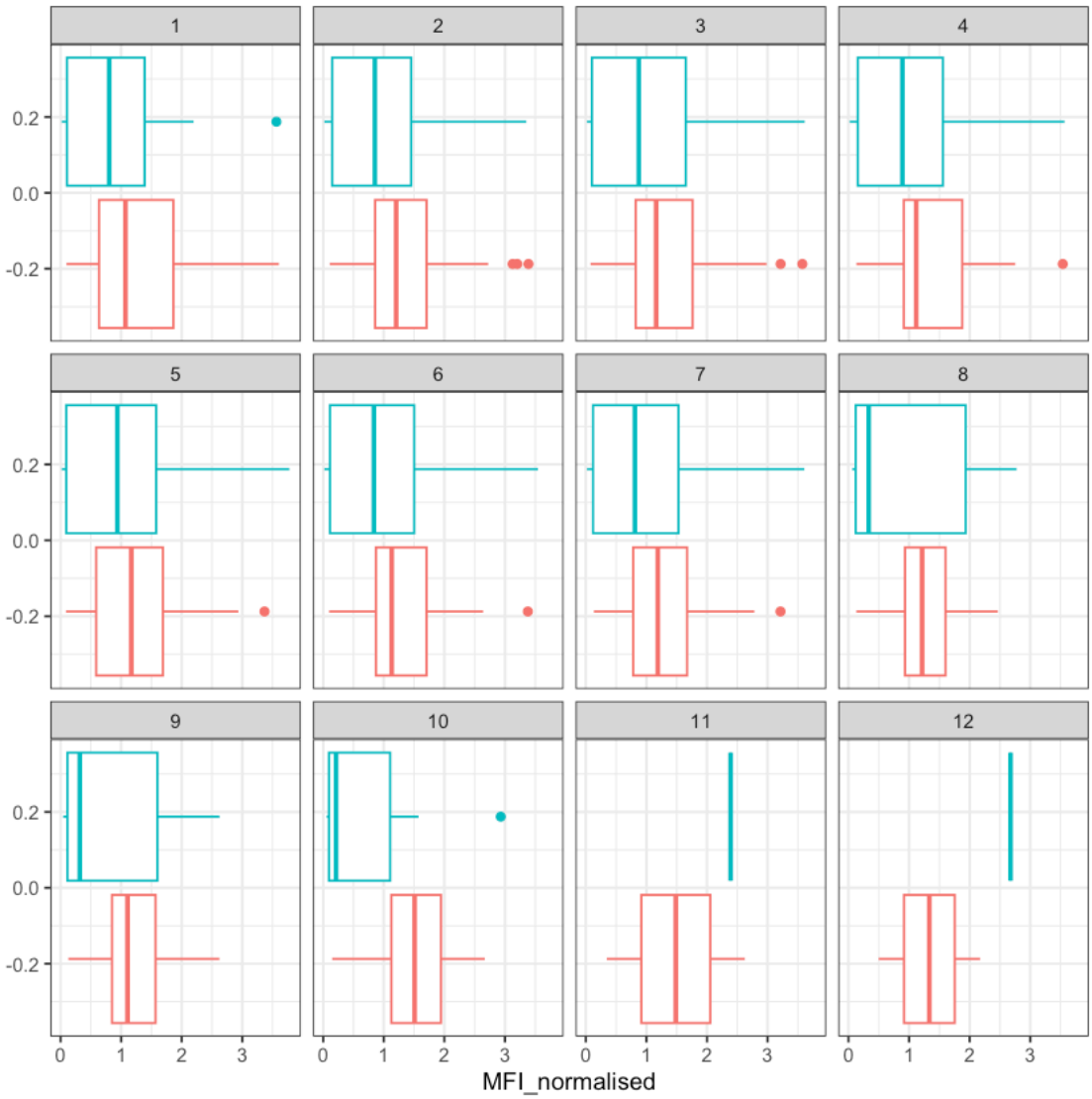
[35]:
```
igg %>% filter(visit != 8) %>%
ggplot() +
aes(MFI_normalised, antigen, col=infancy_vac ) +
geom_boxplot(show.legend = FALSE) +
xlim(0,75) +
facet_wrap(vars(infancy_vac, visit), nrow=2)
```

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



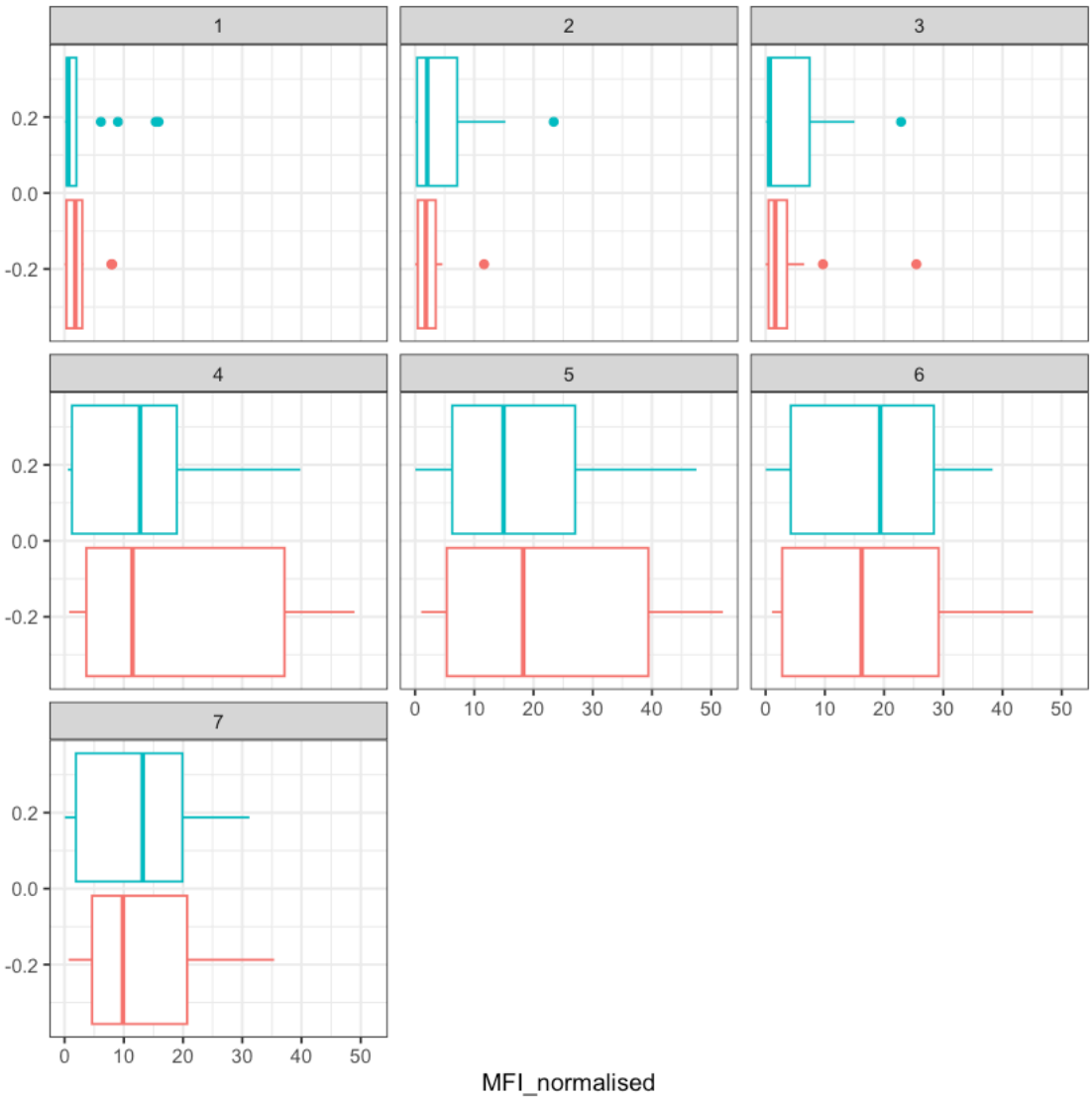
4.7 Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("OVA", that is not in our vaccines) and a clear antigen of interest ("PT", Pertussis Toxin, one of the key virulence factors produced by the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

[39]:
```
filter(igg, antigen=="OVA") %>%
ggplot() +
aes(x=MFI_normalised, col=infancy_vac) +
geom_boxplot(show.legend = FALSE) +
facet_wrap(vars(visit)) +
theme_bw()
```



4.8 and the same for antigen=="FIM2/3"

[40]:
```
filter(igg, antigen=="FIM2/3") %>%
ggplot() +
aes(x=MFI_normalised, col=infancy_vac) +
geom_boxplot(show.legend = FALSE) +
facet_wrap(vars(visit)) +
theme_bw()
```



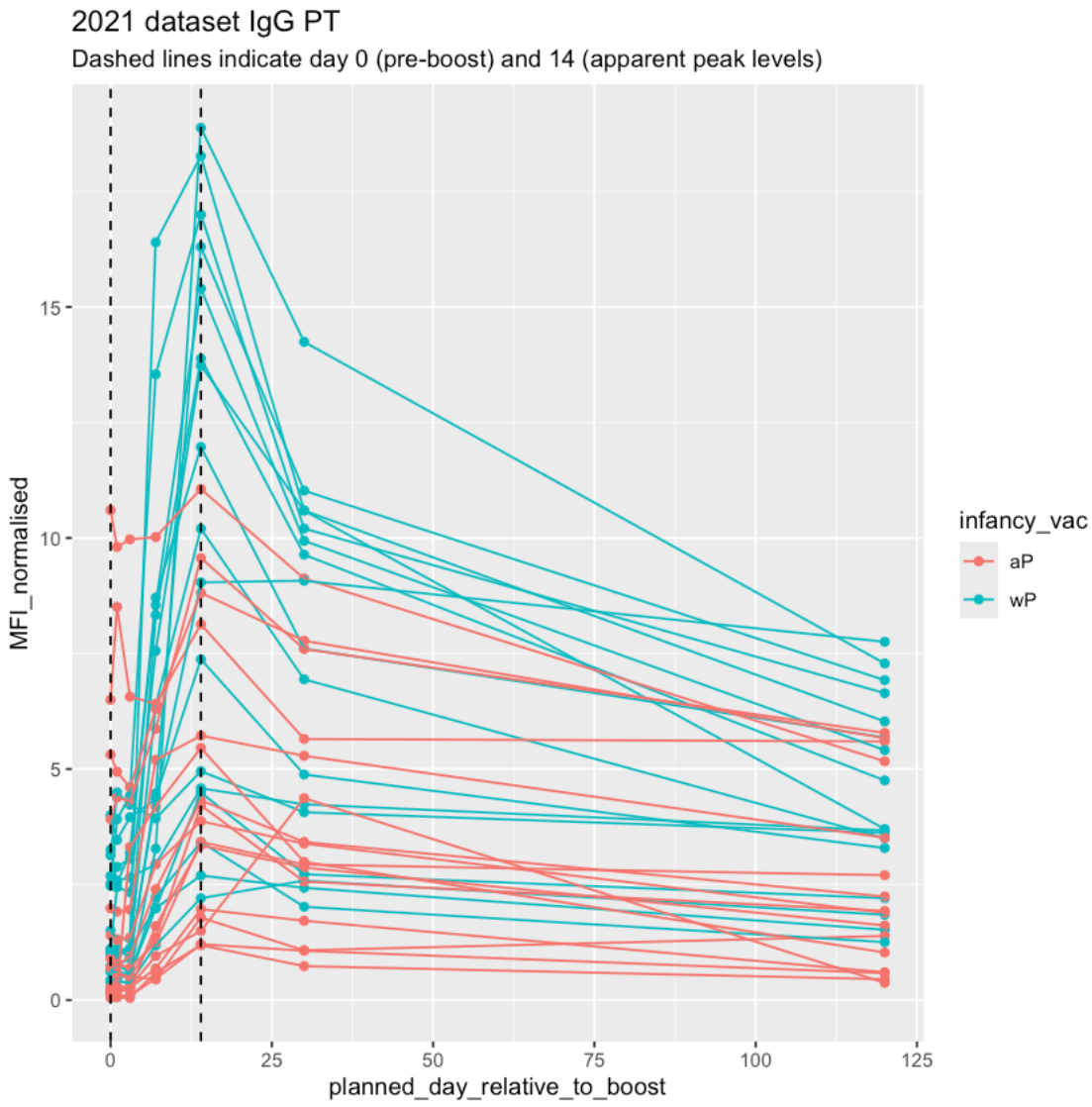
4.9 Q16. What do you notice about these two antigens time courses and the PT data in particular?

PT & FIM2/3 antigens clearly show immunogenic responses through increases in MFI_nomarlized values that peak 5 and then decline.

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

aP and wP for PT & FIM2/3 antigens show magnitude difference in immune response but there is no pattern that separates the two vaccines, hence the difference observed is due to natural variablity in individual responses.

[43]:
```
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)")
```



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

Yes, the trends look similar in both.

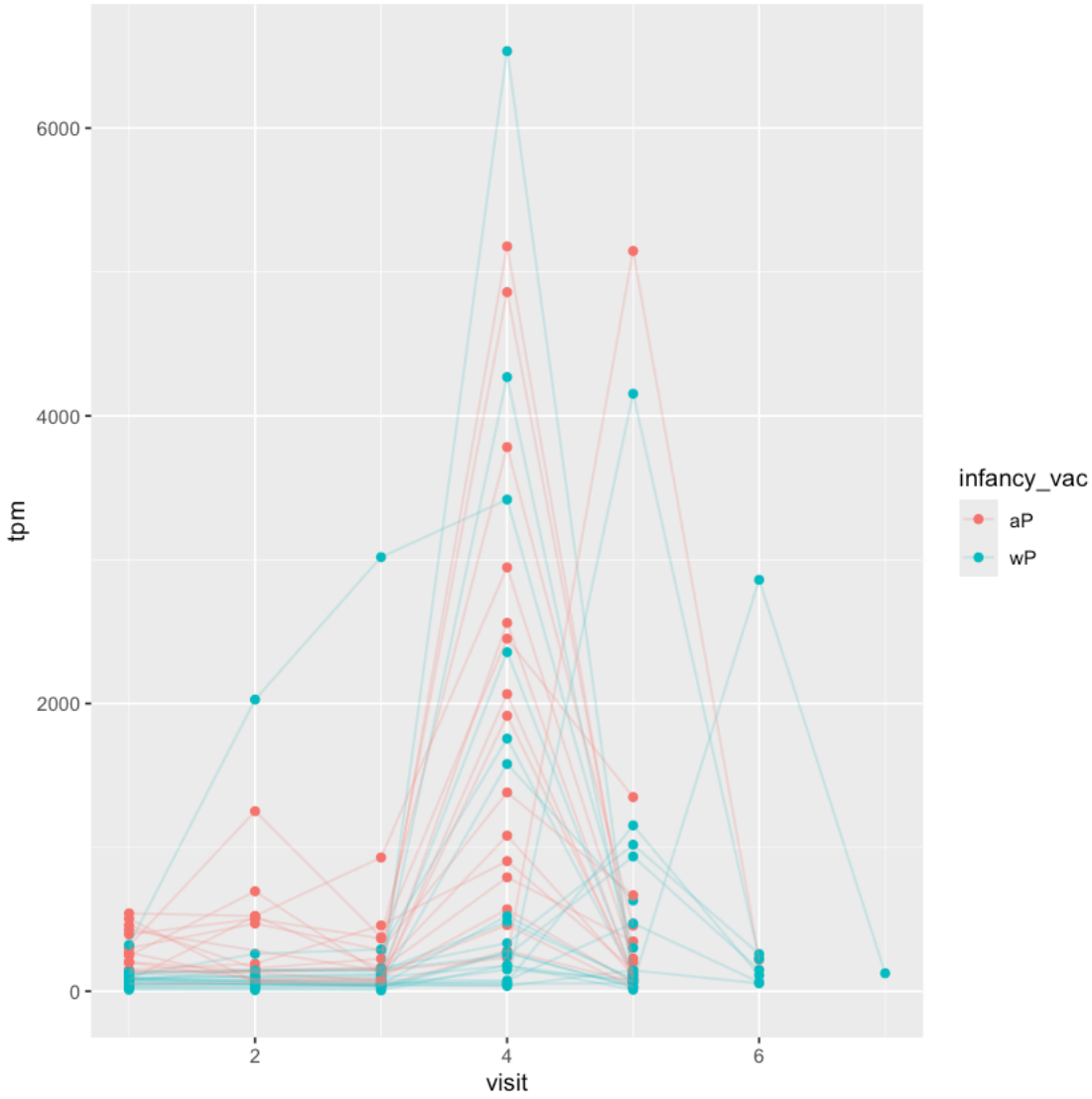
5 5. Obtaining CMI-PB RNASeq data

[44]:
```
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)`
```

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

[46]:
```
ggplot(ssrna) +
  aes(x=visit, y=tpm, color=infancy_vac, group=subject_id) +
  geom_point() +
  geom_line(alpha=0.2)
```



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

Again the peak in 5 demonstrates a significant upregulation post vaccination which is associated to biological activity to immune response, and the area between the rapid peak and followed decline can be indicative of optimal vaccine-induced immune activation

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

Yes it does