

Class 18: Pertussis Mini Project

James Garza PID: A16300772

First we will examine and explore Pertussis case numbers in the US as tracked by the CDC:
<https://www.cdc.gov/pertussis/surv-reporting/cases-by-year.html>

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.

We can use the datapasta package to scape this data from the website into R:

```
cdc <- data.frame(  
  year = c(1922L,1923L,1924L,1925L,  
            1926L,1927L,1928L,1929L,1930L,1931L,  
            1932L,1933L,1934L,1935L,1936L,  
            1937L,1938L,1939L,1940L,1941L,1942L,  
            1943L,1944L,1945L,1946L,1947L,  
            1948L,1949L,1950L,1951L,1952L,  
            1953L,1954L,1955L,1956L,1957L,1958L,  
            1959L,1960L,1961L,1962L,1963L,  
            1964L,1965L,1966L,1967L,1968L,1969L,  
            1970L,1971L,1972L,1973L,1974L,  
            1975L,1976L,1977L,1978L,1979L,1980L,  
            1981L,1982L,1983L,1984L,1985L,  
            1986L,1987L,1988L,1989L,1990L,  
            1991L,1992L,1993L,1994L,1995L,1996L,  
            1997L,1998L,1999L,2000L,2001L,  
            2002L,2003L,2004L,2005L,2006L,2007L,  
            2008L,2009L,2010L,2011L,2012L,  
            2013L,2014L,2015L,2016L,2017L,2018L,  
            2019L,2020L,2021L),  
  cases = c(107473,164191,165418,152003,  
            202210,181411,161799,197371,
```

```

166914,172559,215343,179135,265269,
180518,147237,214652,227319,103188,
183866,222202,191383,191890,109873,
133792,109860,156517,74715,69479,
120718,68687,45030,37129,60886,
62786,31732,28295,32148,40005,
14809,11468,17749,17135,13005,6799,
7717,9718,4810,3285,4249,3036,
3287,1759,2402,1738,1010,2177,2063,
1623,1730,1248,1895,2463,2276,
3589,4195,2823,3450,4157,4570,
2719,4083,6586,4617,5137,7796,6564,
7405,7298,7867,7580,9771,11647,
25827,25616,15632,10454,13278,
16858,27550,18719,48277,28639,32971,
20762,17972,18975,15609,18617,
6124,2116)

)

```

```
## View(cdc)
```

```
head(cdc)
```

```

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

```

I want a plot of cases per year with ggplot

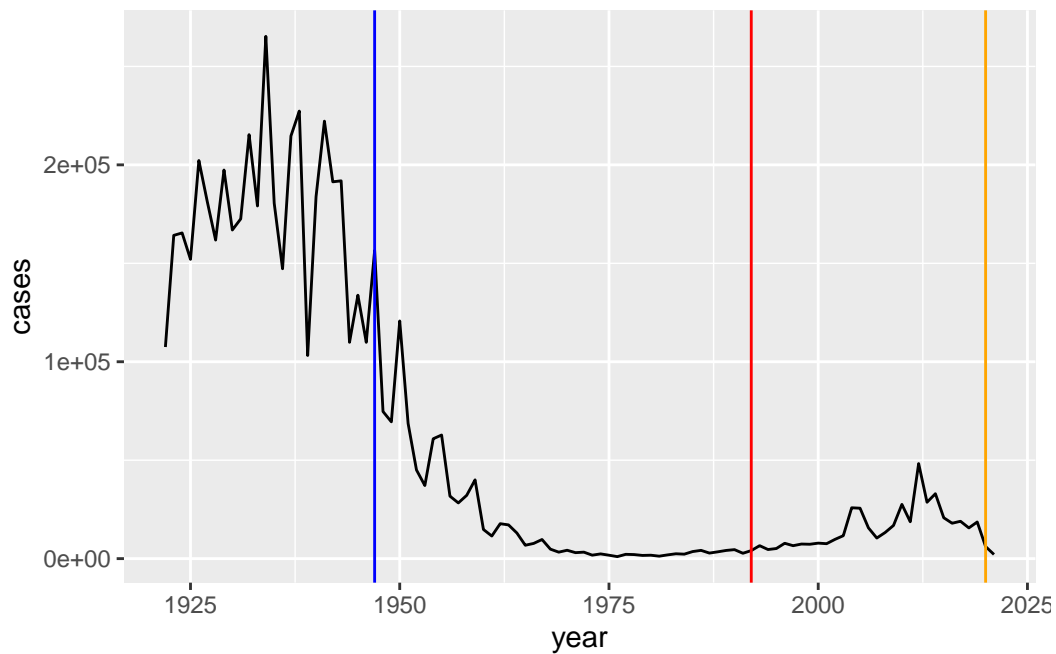
```
library(ggplot2)
```

```

ggplot(cdc) +
  aes(year, cases) +
  geom_line() +
  geom_vline(xintercept=1947, col="blue") +
  geom_vline(xintercept=1992, col="red") +

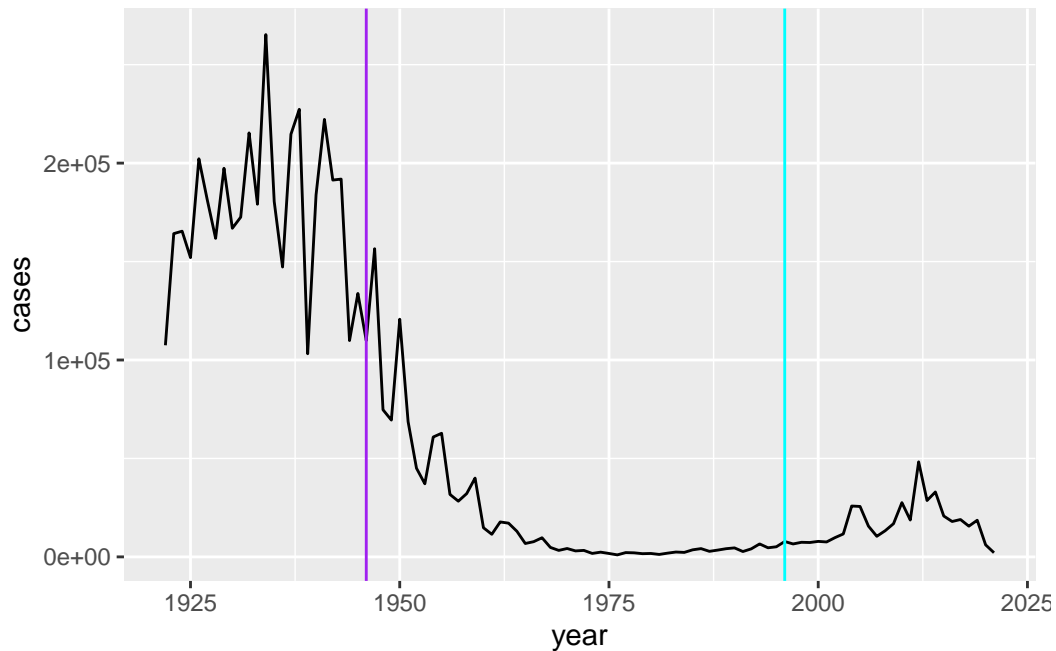
```

```
geom_vline(xintercept=2020, col="orange")
```



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?

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



I noticed that during the switch to the aP vaccine there is the beginning of a small increase to the number of cases but, in between the years there is an overall decrease before

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

10 yr post vaccination efficacy has weakened in the aP vaccine compared to the wP vaccine that was used previously, however there are other theories such as bacterial evolution to the vaccine, vaccine hesitancy, and more sensitive testing.

Exploring CMI-PB data

Access data from the CMI-PB project

This database (like many modern projects) uses an API to return JSON format data.

We will use the R package `jsonlite`.

```
library(jsonlite)

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

Q4. How many aP (newer acellular vaccine) and wP (the older whole cell vaccine) 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$race, subject$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

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

```
table(subject$race, subject$ethnicity)
```

	Hispanic or Latino
American Indian/Alaska Native	0
Asian	0
Black or African American	0
More Than One Race	5
Native Hawaiian or Other Pacific Islander	0
Unknown or Not Reported	10
White	11

	Not Hispanic or Latino	Unknown
American Indian/Alaska Native	1	0
Asian	32	0
Black or African American	2	0
More Than One Race	6	0
Native Hawaiian or Other Pacific Islander	2	0
Unknown or Not Reported	3	2
White	42	2

Side note: working with dates:

We can use the lubridate package to ease the pain of doing math with dates.

```
library(lubridate)
```

Attaching package: 'lubridate'

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

date, intersect, setdiff, union

```
today()
```

```
[1] "2024-03-07"
```

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

Time difference of 8832 days

```
today() - ymd("2002-5-08")
```

Time difference of 7974 days

```
time_length(today() - ymd("2002-5-08"), "years")
```

```
[1] 21.83162
```

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$age <- time_length(today() - ymd(subject$year_of_birth), "years")
subject$age
```

```
[1] 38.17933 56.18070 41.18001 36.18070 33.18001 36.18070 43.17864 39.17864
[9] 28.18070 42.17933 38.17933 42.17933 27.17864 31.17864 35.17864 37.18001
[17] 44.18070 27.17864 30.17933 43.17864 41.18001 39.17864 33.18001 32.18070
[25] 36.18070 41.18001 27.17864 42.17933 27.17864 36.18070 35.17864 27.17864
[33] 34.17933 41.18001 33.18001 27.17864 26.17933 27.17864 39.17864 30.17933
[41] 39.17864 27.17864 26.17933 26.17933 27.17864 26.17933 28.18070 26.17933
[49] 27.17864 27.17864 27.17864 26.17933 26.17933 27.17864 27.17864 27.17864
[57] 28.18070 27.17864 27.17864 27.17864 37.18001 31.17864 29.18001 31.17864
[65] 34.17933 48.18070 52.18070 52.18070 34.17933 26.17933 26.17933 33.18001
[73] 29.18001 29.18001 26.17933 26.17933 36.18070 31.17864 37.18001 32.18070
[81] 31.17864 26.17933 25.18001 27.17864 24.18070 26.17933 24.18070 24.18070
[89] 27.17864 25.18001 26.17933 24.18070 28.18070 25.18001 26.17933 24.18070
[97] 38.17933 31.17864 25.18001 23.17864 21.18001 21.18001 30.17933 35.17864
[105] 30.17933 28.18070 26.17933 29.18001 35.17864 27.17864 28.18070 28.18070
[113] 28.18070 34.17933 22.17933 24.18070 30.17933 26.17933
```

```
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( ap$age ) )
```

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

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

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

```
t.test(ap$age, wp$age, alternative="two.sided")
```

Welch Two Sample t-test

data: ap\$age and wp\$age
t = -12.436, df = 65.411, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-11.950080 -8.643385
sample estimates:
mean of x mean of y
26.27944 36.57618

Since the p value is very small and close to 0 it is significantly different

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

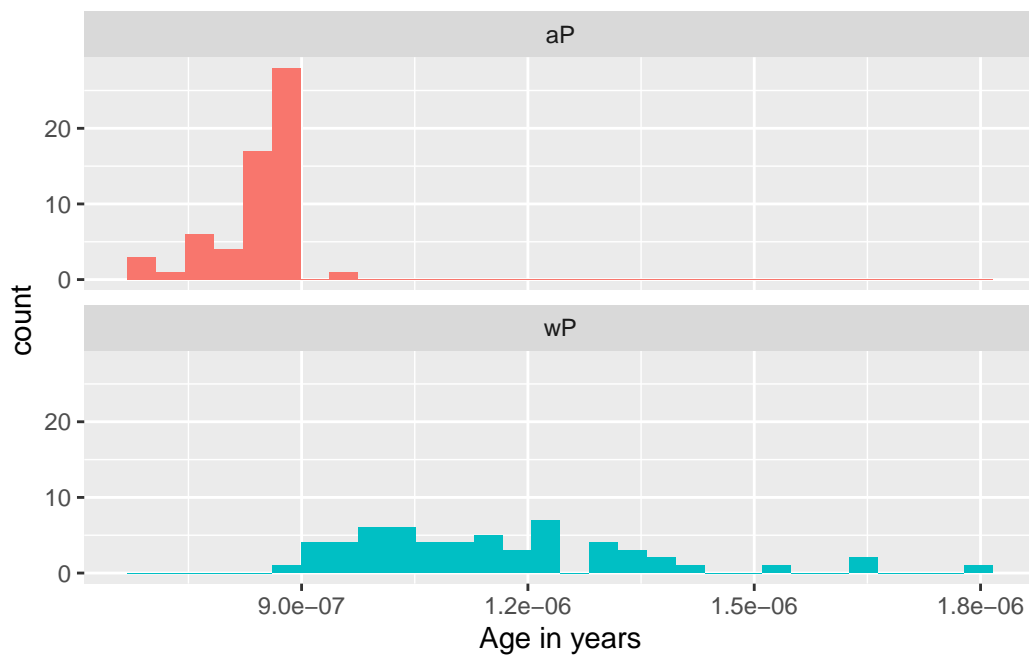
```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

```
[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`.



The groups are different as the distributions of the graphs are very obviously distributed differently.

Get more data from CMI-PB

```
# Complete the API URLs...
specimen <- read_json("http://cmi-pb.org/api/specimen", simplifyVector = TRUE)
titer <- read_json("http://cmi-pb.org/api/v4/plasma_ab_titer", simplifyVector = TRUE)
```

```
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

```
head(titer)
```

	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

We need to **join** these two tables (subject and specimen) to make a single new “meta” table with all our metadata. We will use the `dplyr` function to do this:

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:

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

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

```
dim(meta)
```

```
[1] 939 14
```

```
head(meta)
```

	subject_id	infancy_vac	biological_sex	ethnicity	race
1	1	wP	Female Not Hispanic or Latino	White	
2	1	wP	Female Not Hispanic or Latino	White	
3	1	wP	Female Not Hispanic or Latino	White	
4	1	wP	Female Not Hispanic or Latino	White	
5	1	wP	Female Not Hispanic or Latino	White	
6	1	wP	Female Not Hispanic or Latino	White	

	year_of_birth	date_of_boost	dataset	age	specimen_id
1	1986-01-01	2016-09-12	2020_dataset	38.17933	1
2	1986-01-01	2016-09-12	2020_dataset	38.17933	2
3	1986-01-01	2016-09-12	2020_dataset	38.17933	3
4	1986-01-01	2016-09-12	2020_dataset	38.17933	4
5	1986-01-01	2016-09-12	2020_dataset	38.17933	5
6	1986-01-01	2016-09-12	2020_dataset	38.17933	6

	actual_day_relative_to_boost	planned_day_relative_to_boost	specimen_type
1	-3	0	Blood
2	1	1	Blood

3	3	3	Blood
4	7	7	Blood
5	11	14	Blood
6	32	30	Blood

	visit
1	1
2	2
3	3
4	4
5	5
6	6

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

One more `inner_join` to add all our metadata in `meta` on to our `ab_data` table:

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

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

```
head(abdata)
```

	specimen_id	isotype	is_antigen_specific	antigen	MFI	MFI_normalised
1	1	IgE	FALSE	Total	1110.21154	2.493425
2	1	IgE	FALSE	Total	2708.91616	2.493425
3	1	IgG	TRUE	PT	68.56614	3.736992
4	1	IgG	TRUE	PRN	332.12718	2.602350
5	1	IgG	TRUE	FHA	1887.12263	34.050956
6	1	IgE	TRUE	ACT	0.10000	1.000000

	unit	lower_limit_of_detection	subject_id	infancy_vac	biological_sex
1	UG/ML	2.096133	1	wP	Female
2	IU/ML	29.170000	1	wP	Female
3	IU/ML	0.530000	1	wP	Female
4	IU/ML	6.205949	1	wP	Female
5	IU/ML	4.679535	1	wP	Female
6	IU/ML	2.816431	1	wP	Female

	ethnicity	race	year_of_birth	date_of_boost	dataset
1	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
2	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
3	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset

```

4 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
5 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
6 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
  age actual_day_relative_to_boost planned_day_relative_to_boost
1 38.17933                -3                0
2 38.17933                -3                0
3 38.17933                -3                0
4 38.17933                -3                0
5 38.17933                -3                0
6 38.17933                -3                0
  specimen_type visit
1      Blood      1
2      Blood      1
3      Blood      1
4      Blood      1
5      Blood      1
6      Blood      1

```

```
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

```

Examine IgG Ab titer levels

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

	specimen_id	isotype	is_antigen_specific	antigen	MFI	MFI_normalised
1	1	IgG	TRUE	PT	68.56614	3.736992
2	1	IgG	TRUE	PRN	332.12718	2.602350
3	1	IgG	TRUE	FHA	1887.12263	34.050956
4	19	IgG	TRUE	PT	20.11607	1.096366
5	19	IgG	TRUE	PRN	976.67419	7.652635
6	19	IgG	TRUE	FHA	60.76626	1.096457

	unit	lower_limit_of_detection	subject_id	infancy_vac	biological_sex
1	IU/ML	0.530000	1	wP	Female
2	IU/ML	6.205949	1	wP	Female
3	IU/ML	4.679535	1	wP	Female
4	IU/ML	0.530000	3	wP	Female
5	IU/ML	6.205949	3	wP	Female
6	IU/ML	4.679535	3	wP	Female

	ethnicity	race	year_of_birth	date_of_boost	dataset
1	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
2	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
3	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
4	Unknown	White	1983-01-01	2016-10-10	2020_dataset
5	Unknown	White	1983-01-01	2016-10-10	2020_dataset
6	Unknown	White	1983-01-01	2016-10-10	2020_dataset

	age	actual_day_relative_to_boost	planned_day_relative_to_boost
1	38.17933	-3	0
2	38.17933	-3	0
3	38.17933	-3	0
4	41.18001	-3	0
5	41.18001	-3	0
6	41.18001	-3	0

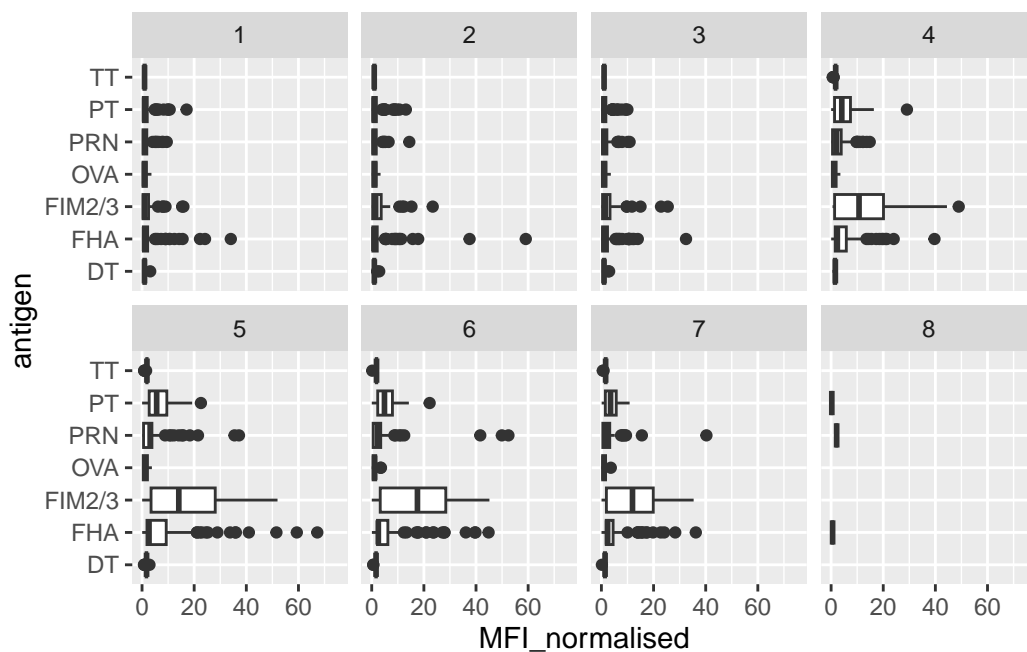
	specimen_type	visit
1	Blood	1
2	Blood	1
3	Blood	1
4	Blood	1
5	Blood	1
6	Blood	1

Q13. Complete the following code to make a summary boxplot of Ab titer levels

(MFI) for all antigens:

```
ggplot(igg) +  
  aes(MFI_normalised, antigen) +  
  geom_boxplot() +  
  xlim(0,75) +  
  facet_wrap(vars(visit), nrow=2)
```

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

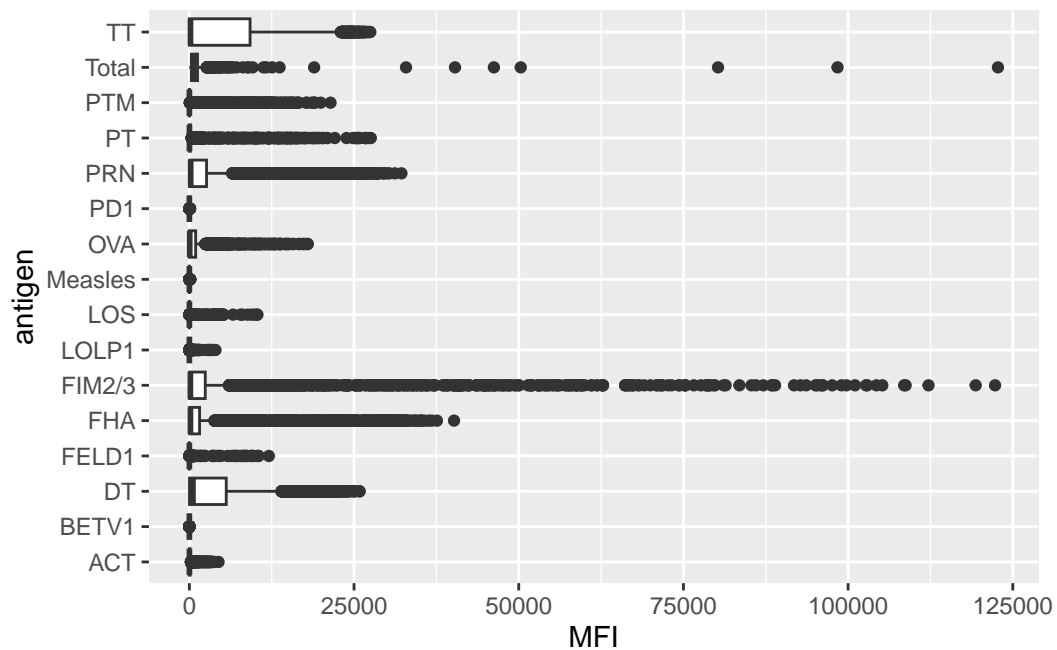


Q14. What antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others?

FHA, PRN PT and FIM2/3 both show differences over time compared to the other antigens because these antigens were included in the vaccine so the MFI is higher.

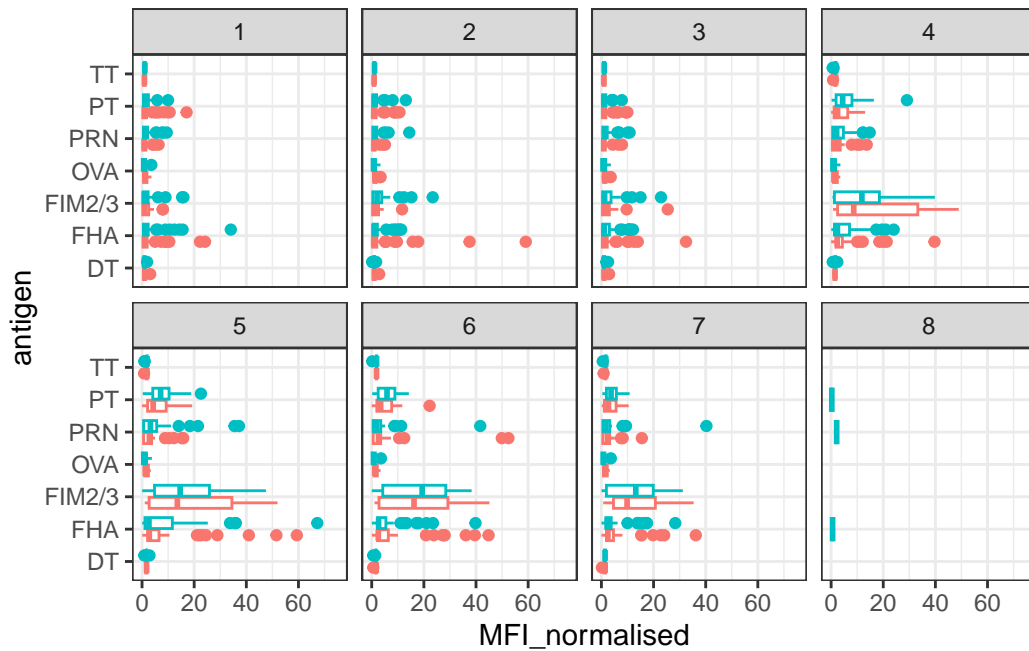
```
ggplot(abdata) +  
  aes(MFI, antigen) +  
  geom_boxplot()
```

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

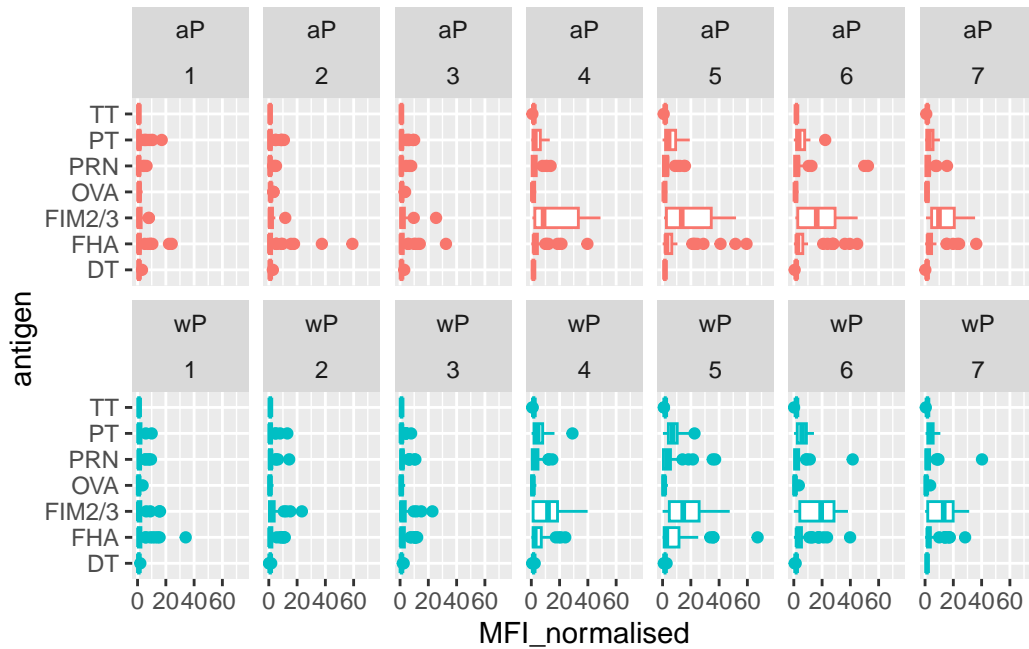


```

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)

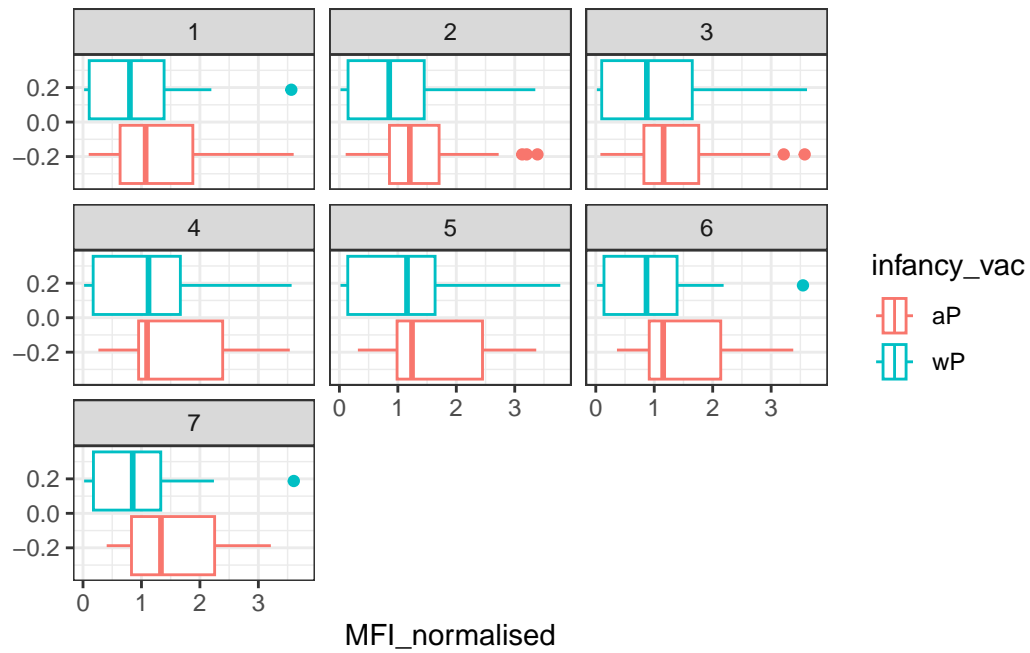
```

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

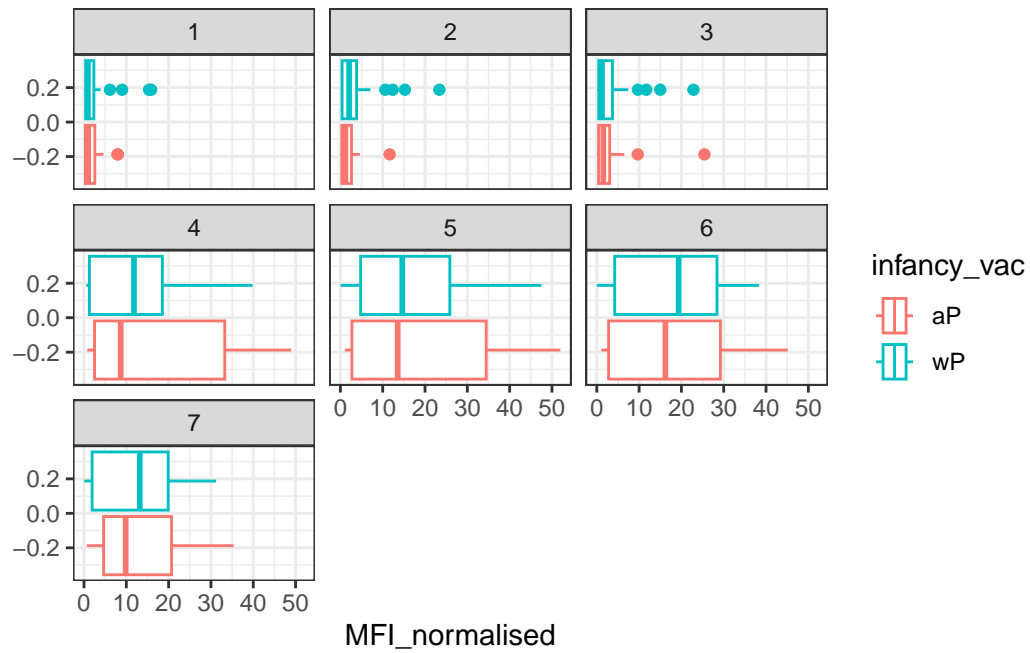


Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a “control” antigen (“OVA”, that is not in our vaccines) and a clear antigen of interest (“PT”, Pertussis Toxin, one of the key virulence factors produced by the bacterium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

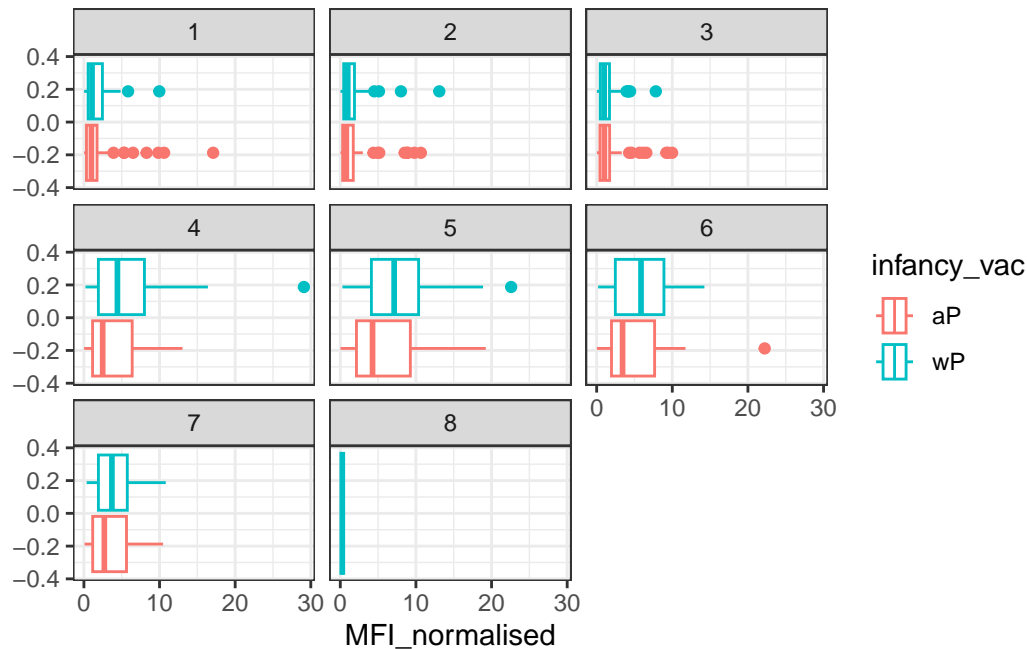
```
filter(igg, antigen=="OVA") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = TRUE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



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



```
filter(igg, antigen=="PT") %>%
  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?

PT increases over time but then declines a lot but the other two increase more slowly over time and only slightly decrease.

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

They appear to be the same as there is no clear significant difference between the two for the boxplot. But for the lineplot, there is a clear difference between wP and aP as around day 14 there is a large increase in wP but not aP.

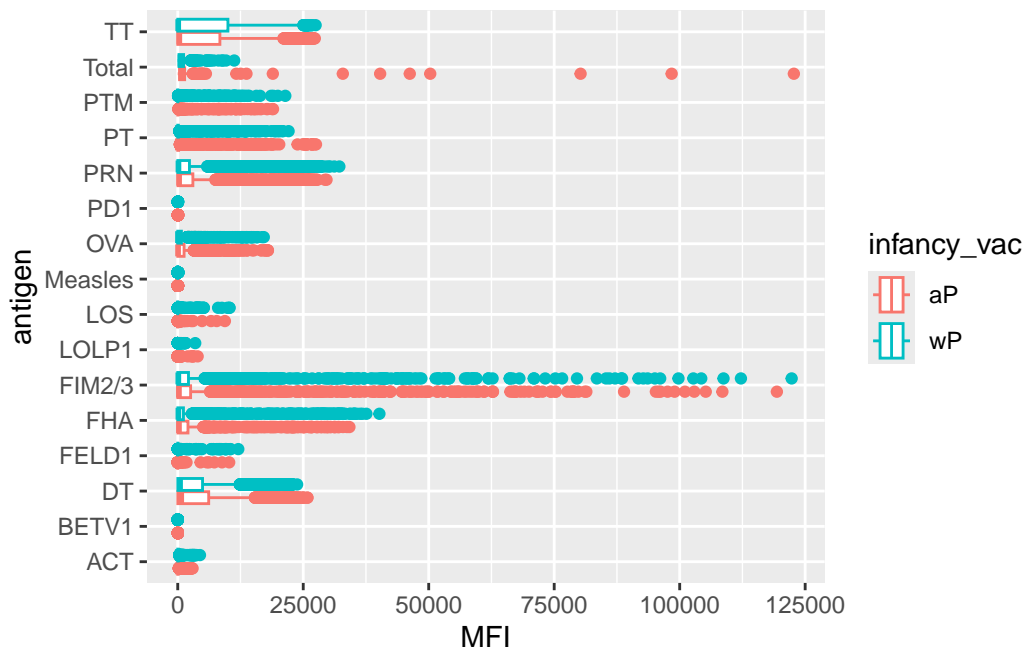
Why are certain antigens and not others very variable in their detected levels here?

The vaccine only contains certain antigens

Can you facet or even just color by infancy_vac?

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

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



There are potentially some differences here but in general it is hard to tell with this whole dataset overview...

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

```
2020_dataset 2021_dataset 2022_dataset
      31520         8085         2170
```

Let's focus in on jsut the 2021 dataset.

```
abdata.21 <- filter(abdata, dataset == "2021_dataset")
table(abdata.21$dataset)
```

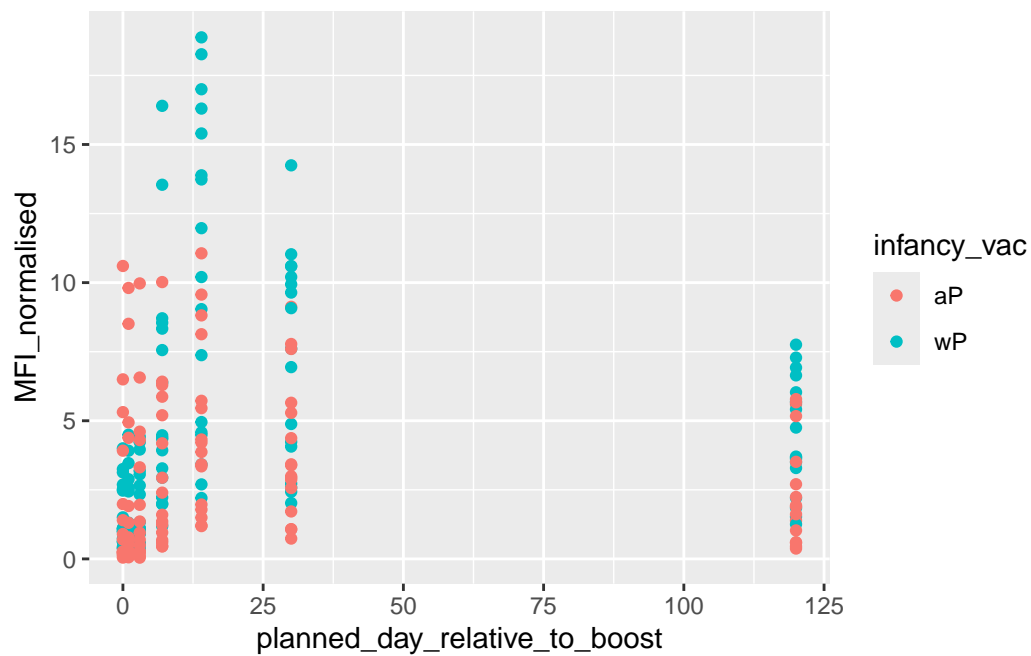
```
2021_dataset
      8085
```

Focus on PT antigen for IgG levels

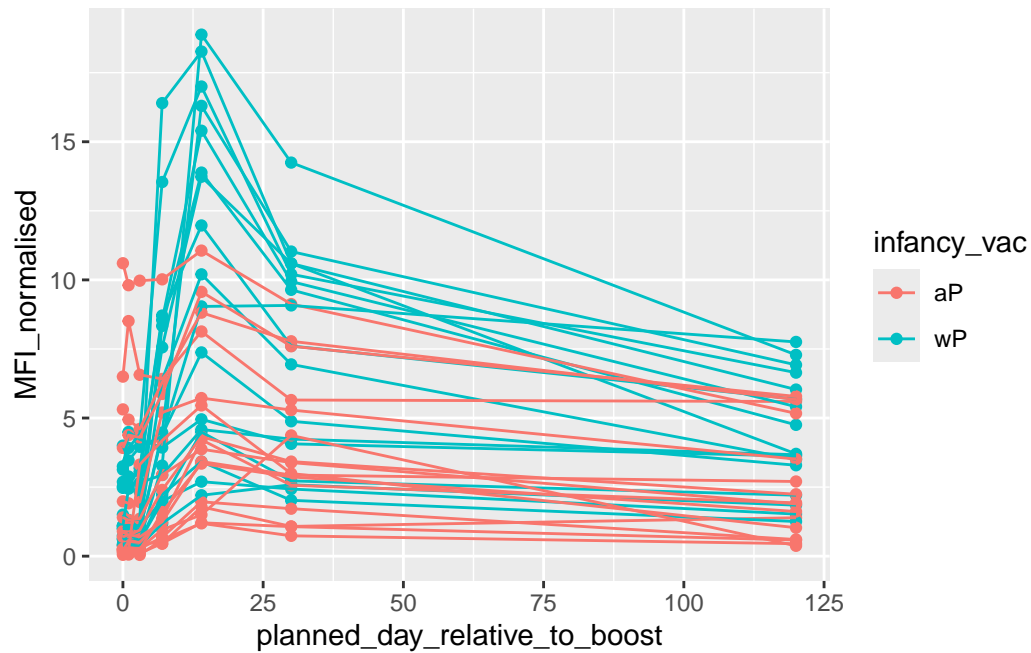
```
pt.21 <- filter(abdata.21, isotype == "IgG", antigen == "PT")
```

plot of days (time)

```
ggplot(pt.21) +
  aes(x=planned_day_relative_to_boost,
      y=MFI_normalised,
      col=infancy_vac) +
  geom_point()
```



```
ggplot(pt.21) +
  aes(x=planned_day_relative_to_boost,
      y=MFI_normalised,
      col=infancy_vac, group=subject_id) +
  geom_point() +
  geom_line()
```



```
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)")
```


2021 dataset IgG PT

Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)

