

# Class 18

Josefina O'Toole (PID: A16978557)

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.

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.

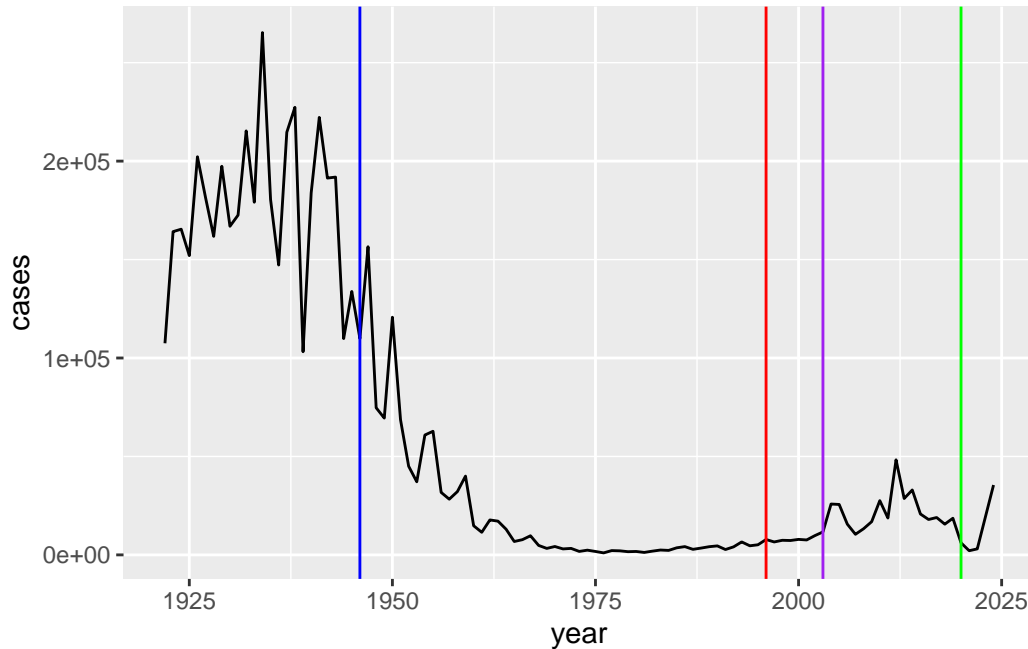
```
head(cdc)
```

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

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="green") +
  geom_vline(xintercept = 2003, col="purple")
```



There were high cases 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?

There is a lag, a waning of immunity if you have the aP vaccine. There is an initial decrease in cases and then a resurgence. The immunity generated by the acellular pertussis vaccine is not as long-lasting as that from the whole-cell vaccine.

What is different about the immune response to infection if you had an older wP vaccine v.s. 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 = 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

How many individual “subjects” are in this dataset?

```
nrow(subject)
```

```
[1] 172
```

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

```
library(lubridate)
```

Attaching package: 'lubridate'

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

date, intersect, setdiff, union

```
today()
```

```
[1] "2025-03-09"
```

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

Time difference of 9199 days

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

```
[1] 25.18549
```

```
# Use todays date to calculate age in days
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.
22	26	27	27	28	34

```
# wP
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

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?

- (i) 36
- (ii) 27
- (iii) They are significantly different.

```
x <- t.test(time_length( wp$age, "years" ),
            time_length( ap$age, "years" ))
x$p.value
```

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

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)
```

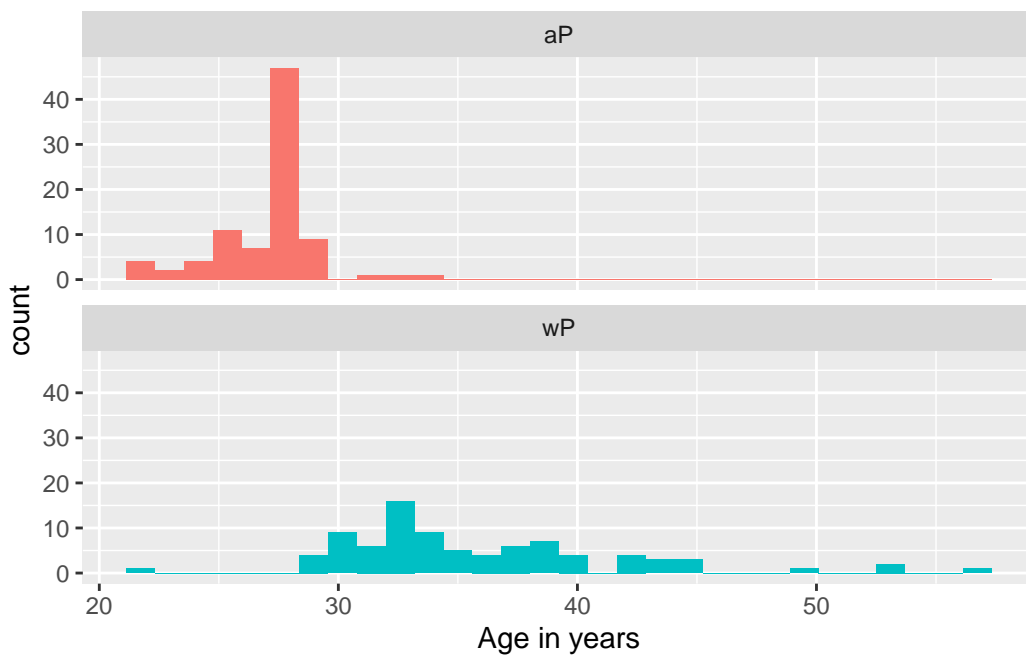
```
[1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481
```

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

I 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`.



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:

Obtain more data from CMI-PB...

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

```
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(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.

Q9.2

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.

```
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	specimen_id
1	1986-01-01	2016-09-12	2020_dataset	14312 days	1
2	1986-01-01	2016-09-12	2020_dataset	14312 days	2
3	1986-01-01	2016-09-12	2020_dataset	14312 days	3
4	1986-01-01	2016-09-12	2020_dataset	14312 days	4
5	1986-01-01	2016-09-12	2020_dataset	14312 days	5
6	1986-01-01	2016-09-12	2020_dataset	14312 days	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   9
```

```
dim(specimen)
```

```
[1] 1503   6
```

```
dim(meta)
```

```
[1] 1503  14
```

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	specimen_id	
1	1986-01-01	2016-09-12	2020_dataset	14312 days		1
2	1986-01-01	2016-09-12	2020_dataset	14312 days		1
3	1986-01-01	2016-09-12	2020_dataset	14312 days		1
4	1986-01-01	2016-09-12	2020_dataset	14312 days		1

```

5    1986-01-01    2016-09-12 2020_dataset 14312 days          1
6    1986-01-01    2016-09-12 2020_dataset 14312 days          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

```

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?

The number of rows on the most “recent” dataset is half as large as the number of rows from 2020 but twice as large as the number of rows for 2021 and 2022.

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

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

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

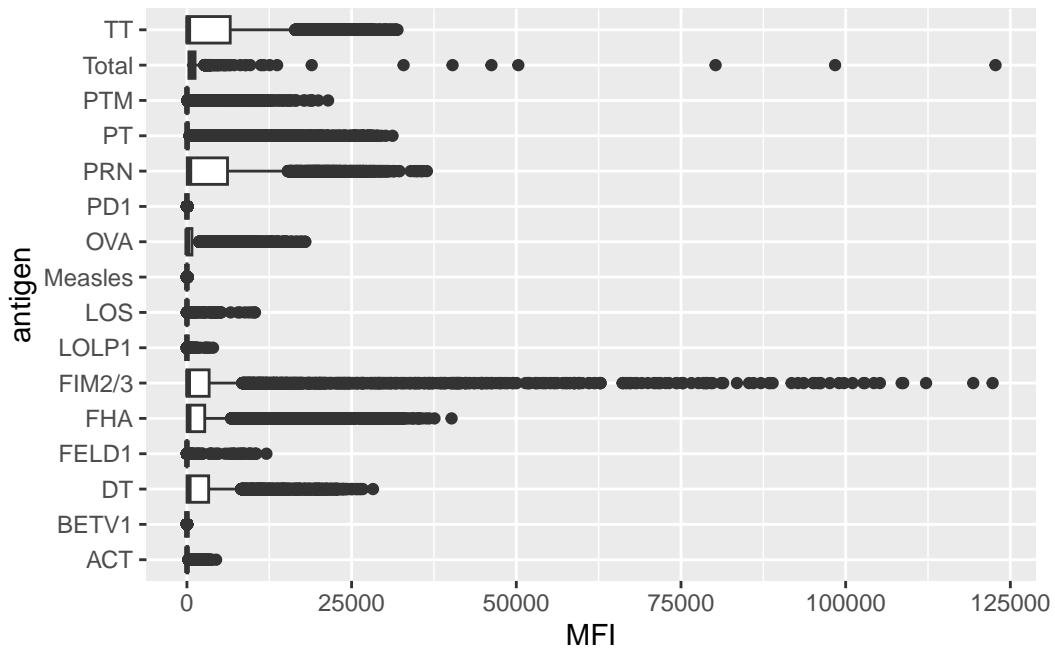
```

      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
```

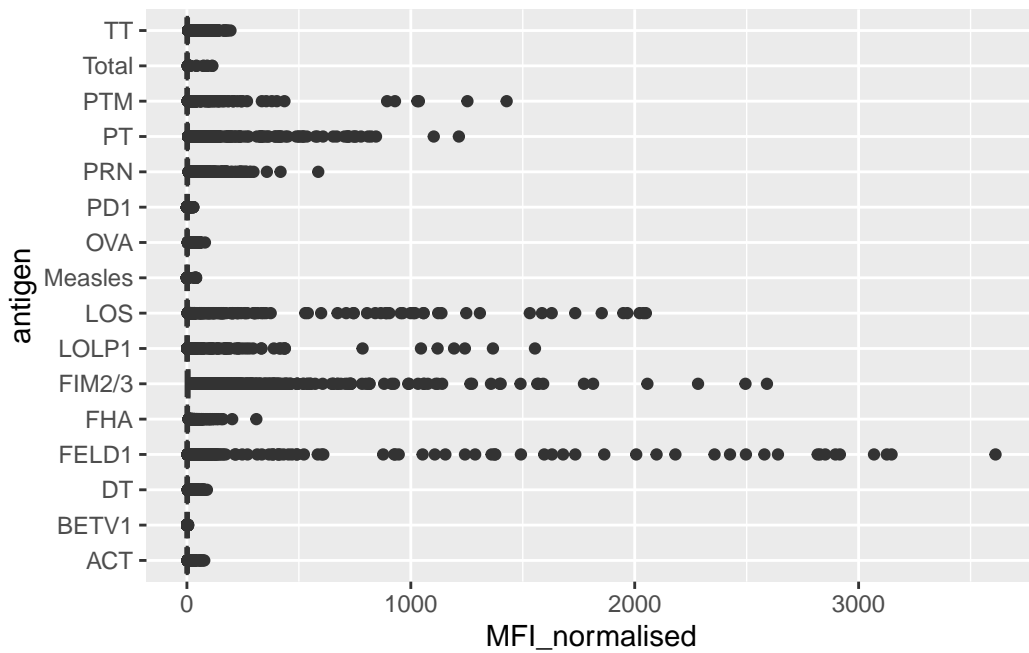
I want a plot of antigen levels across the whole dataset.

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

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

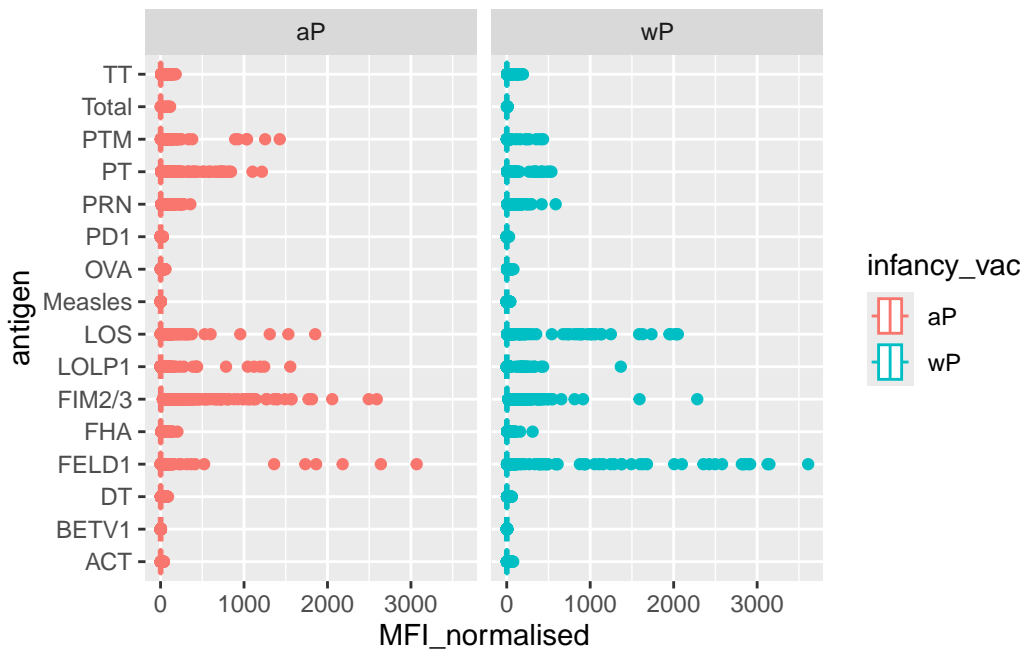


```
ggplot(abdata) +
  aes(MFI_normalised, antigen) +
  geom_boxplot()
```



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

```
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.

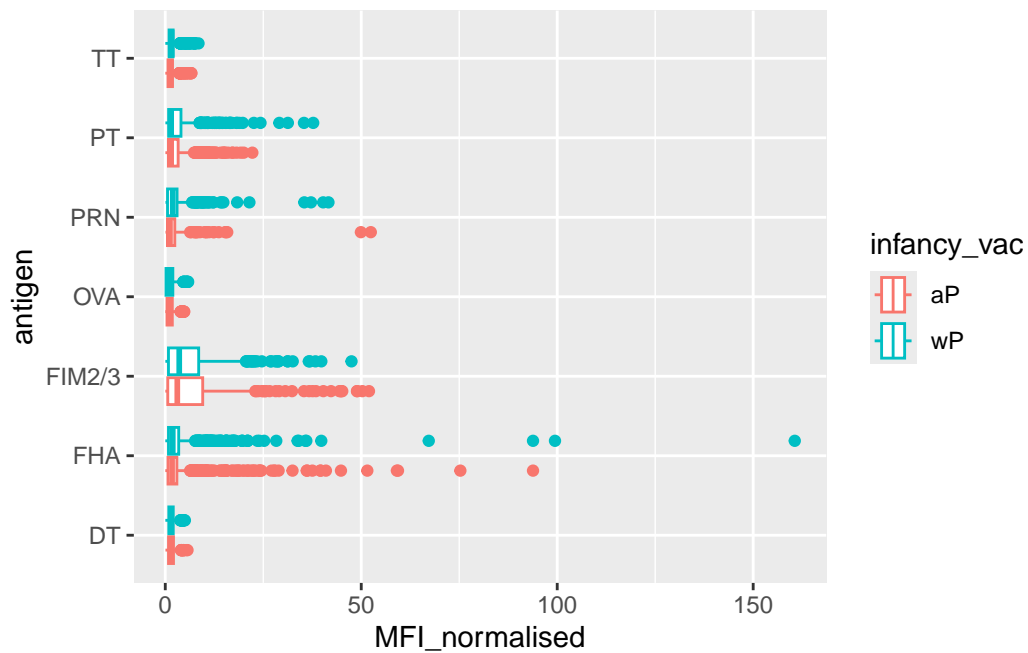
```
igg <- abdata %>% filter(isotype == "IgG")
head(igg)
```

	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

	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	14312 days	1		
2	1986-01-01	2016-09-12	2020_dataset	14312 days	1		
3	1986-01-01	2016-09-12	2020_dataset	14312 days	1		
4	1986-01-01	2016-09-12	2020_dataset	14312 days	2		
5	1986-01-01	2016-09-12	2020_dataset	14312 days	2		
6	1986-01-01	2016-09-12	2020_dataset	14312 days	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:

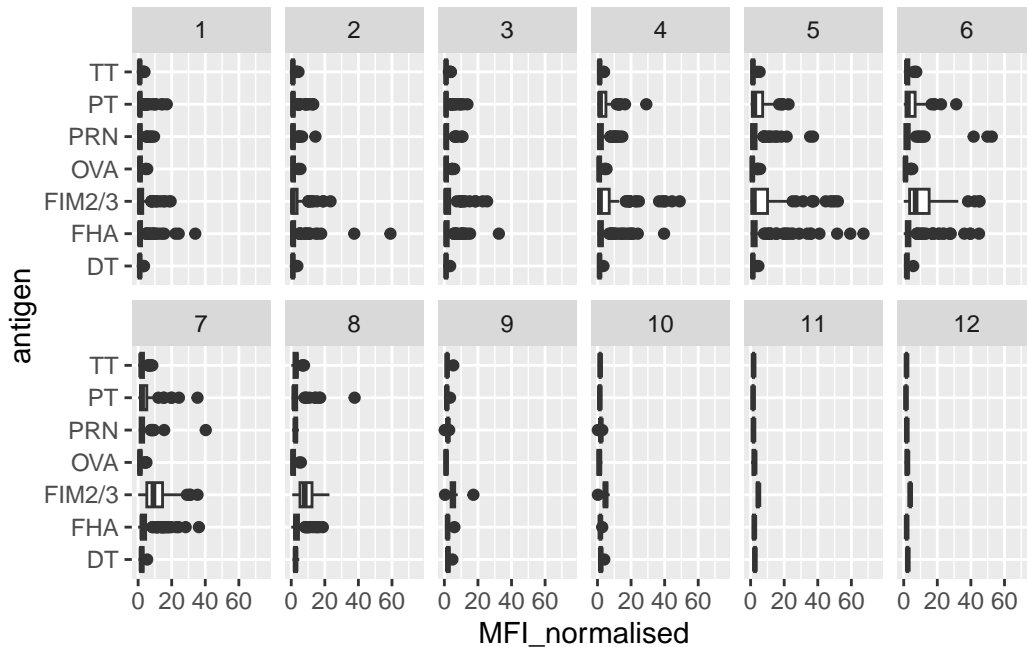
```
ggplot(igg) +
  aes(MFI_normalised, antigen, col=infancy_vac) +
  geom_boxplot()
```



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?

The Fim2/3 antigen, a component of acellular pertussis vaccines, refers to the fimbriae (Fim) types 2 and 3 produced by *Bordetella pertussis*, the bacteria causing whooping cough, which are important for bacterial adhesion and are used as vaccine antigens.

Digging in further to look at the 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
abdata.21 %>%
  filter(isotype == "IgG", antigen == "PT") %>%

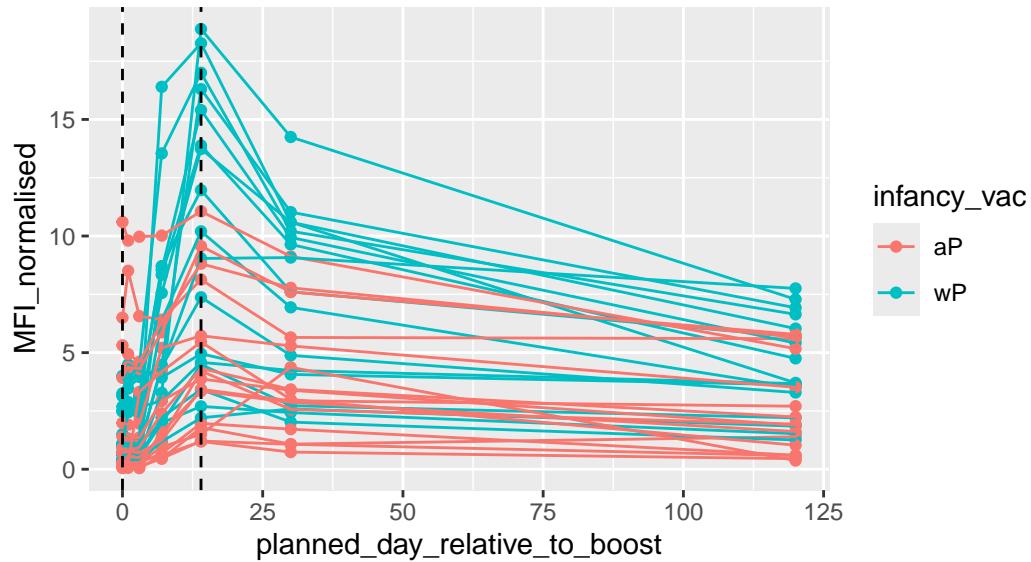
  #Plot and color by infancy_vac (wP vs aP)
ggplot() +
  aes(x=planned_day_relative_to_boost,
       y=MFI_normalised,
       col=infancy_vac,
       group=subject_id) +
  geom_point() +
```



```
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)



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