

AUTHOR
Lisa Chen A17082974

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

Q1 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.

[illegible]

```
202210,181411,161799,
166914,172559,215343,
180518,147237,214652,
183866,222202,191383,
133792,109860,156517,
120718,68687,45030,37
62786,31732,28295,321
14809,11468,17749,171
7717,9718,4810,3285,4
3287,1759,2402,1738,1
1623,1730,1248,1895,2
3589,4195,2823,3450,4
2719,4083,6586,4617,5
7405,7298,7867,7580,9
25827,25616,15632,104
16858,27550,18719,482
20762,17972,18975,156
6124,2116)
```

```
)
```

```
View(cdc)
```

```
head(cdc)
```

	year	No. Reported. Pertussis. Cases
1	1922	107473
2	1923	164191
3	1924	165418
4	1925	152003
5	1926	202210
6	1927	181411

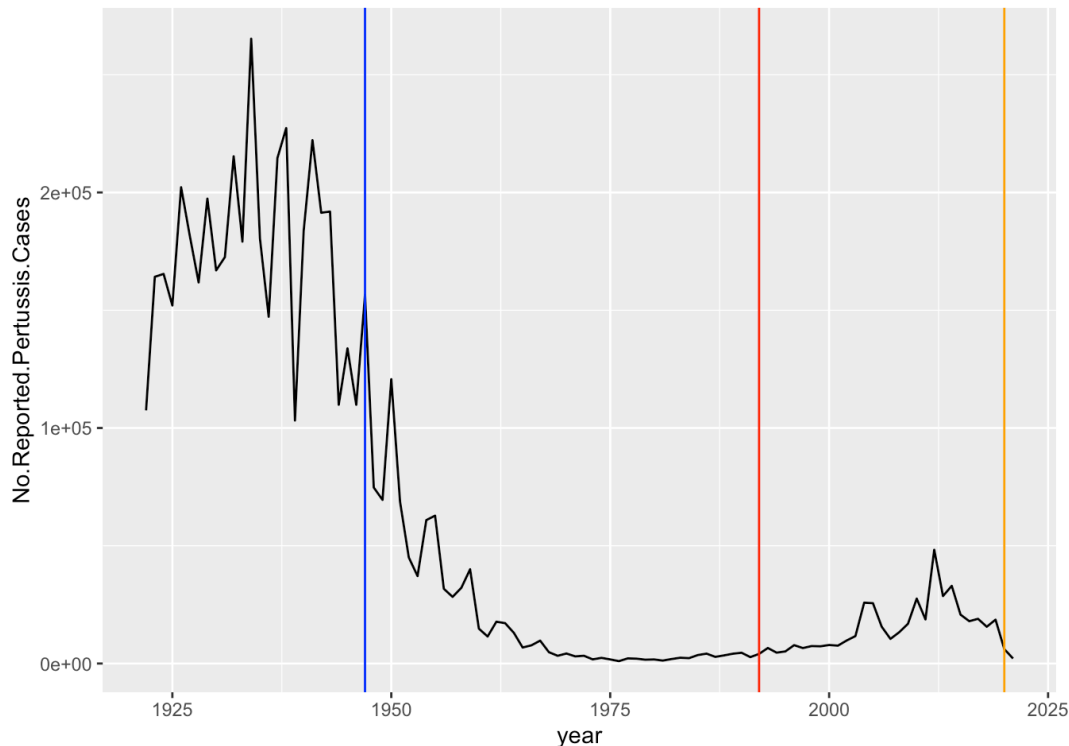
I want a plot of cases per year with ggplot

```
library(ggplot2)
```

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, No.Reported.Pertussis.Cases) +  
  geom_line() +  
  geom_vline(xintercept = 1947, col="blue") +  
  geom_vline(xintercept = 1992, col="red") +  
  geom_vline(xintercept = 2020, col="orange")
```



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

There is an increase most likely due to COVID and comorbidities that could contribute to decreased immunity. There could also be 1) more sensitive PCR-based testing, 2) vaccination hesitancy 3) bacterial evolution (escape from vaccine immunity), 4) waning of immunity in adolescents originally primed as infants with the newer aP vaccine as compared to the older wP vaccine.

Access data from the CMI-PB project

This database(like many modern project) uses an APi to return JSON format data.

We will use the R package jsonlite

```
library(jsonlite)

subject <- read_json("http://cmi-pb.org/api/subject",
                      simplifyVector = TRUE)

head(subject)
```

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

	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

q4How many wP(the older whole-cell vaccine) individuals and aP (newer acellular vaccine) individuals are in this dataset?

```
table(subject$infancy_vac)
```

```
aP wP
60 58
```

Q5. What is the number of individuals by biological sex and race

```
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$race, subject$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

```
subject$year_of_birth
```

```
[1] "1986-01-01" "1968-01-01" "1983-01-01" "1988-01-01"
"1991-01-01"
[6] "1988-01-01" "1981-01-01" "1985-01-01" "1996-01-01"
"1982-01-01"
[11] "1986-01-01" "1982-01-01" "1997-01-01" "1993-01-01"
"1989-01-01"
[16] "1987-01-01" "1980-01-01" "1997-01-01" "1994-01-01"
"1981-01-01"
[21] "1983-01-01" "1985-01-01" "1991-01-01" "1992-01-01"
"1988-01-01"
[26] "1983-01-01" "1997-01-01" "1982-01-01" "1997-01-01"
"1988-01-01"
[31] "1989-01-01" "1997-01-01" "1990-01-01" "1983-01-01"
"1991-01-01"
[36] "1997-01-01" "1998-01-01" "1997-01-01" "1985-01-01"
"1994-01-01"
[41] "1985-01-01" "1997-01-01" "1998-01-01" "1998-01-01"
"1997-01-01"
[46] "1998-01-01" "1996-01-01" "1998-01-01" "1997-01-01"
"1997-01-01"
[51] "1997-01-01" "1998-01-01" "1998-01-01" "1997-01-01"
```

```

"1997-01-01"
[56] "1997-01-01" "1996-01-01" "1997-01-01" "1997-01-01"
"1997-01-01"
[61] "1987-01-01" "1993-01-01" "1995-01-01" "1993-01-01"
"1990-01-01"
[66] "1976-01-01" "1972-01-01" "1972-01-01" "1990-01-01"
"1998-01-01"
[71] "1998-01-01" "1991-01-01" "1995-01-01" "1995-01-01"
"1998-01-01"
[76] "1998-01-01" "1988-01-01" "1993-01-01" "1987-01-01"
"1992-01-01"
[81] "1993-01-01" "1998-01-01" "1999-01-01" "1997-01-01"
"2000-01-01"
[86] "1998-01-01" "2000-01-01" "2000-01-01" "1997-01-01"
"1999-01-01"
[91] "1998-01-01" "2000-01-01" "1996-01-01" "1999-01-01"
"1998-01-01"
[96] "2000-01-01" "1986-01-01" "1993-01-01" "1999-01-01"
"2001-01-01"
[101] "2003-01-01" "2003-01-01" "1994-01-01" "1989-01-01"
"1994-01-01"
[106] "1996-01-01" "1998-01-01" "1995-01-01" "1989-01-01"
"1997-01-01"
[111] "1996-01-01" "1996-01-01" "1996-01-01" "1990-01-01"
"2002-01-01"
[116] "2000-01-01" "1994-01-01" "1998-01-01"

```

#Side-Note: Working with dates

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

Q7 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

```
today()
```

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

```
today() - ymd("2002-01-01")
```

Time difference of 8101 days

```
today() - mdy("5-15-2002")
```

Time difference of 7967 days

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
```

So what is the age of everyone on our dataset.

```
time_length(today() - mdy("5-15-2002"), "years")
```

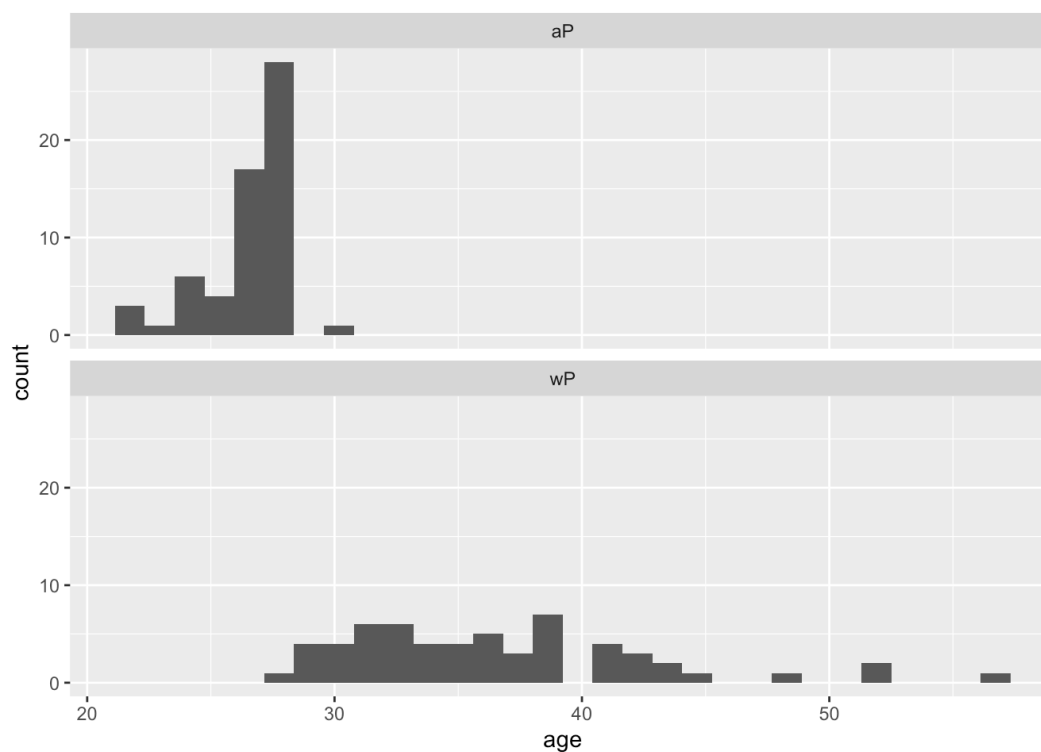
```
[1] 21.81246
```

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

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

```
library(ggplot2)
ggplot(subject) +
  aes(age) +
    facet_wrap(vars(infancy_vac), nrow=2) +
    geom_histogram()
```

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.



##Get more data from CMi-PB

```
specimen <- read_json("http://cmi-pb.org/api/specimen", simplify = FALSE)
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

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:

we need to **join** these two tables(subject and specimen) to make a single new "meta" table with all of our metadata

```
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

```
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	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

```

Now we can read some of the other data from CMI-PB

```

ab_titer <- read_json("http://cmi-pb.org/api/v4/plasma_ab_titer")
head(ab_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

```

One more 'inner_join()' to add all our metadata in 'meta' on to our 'ab_data' table:

Q.10

```
abdata <- inner_join(ab_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

Q11

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

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

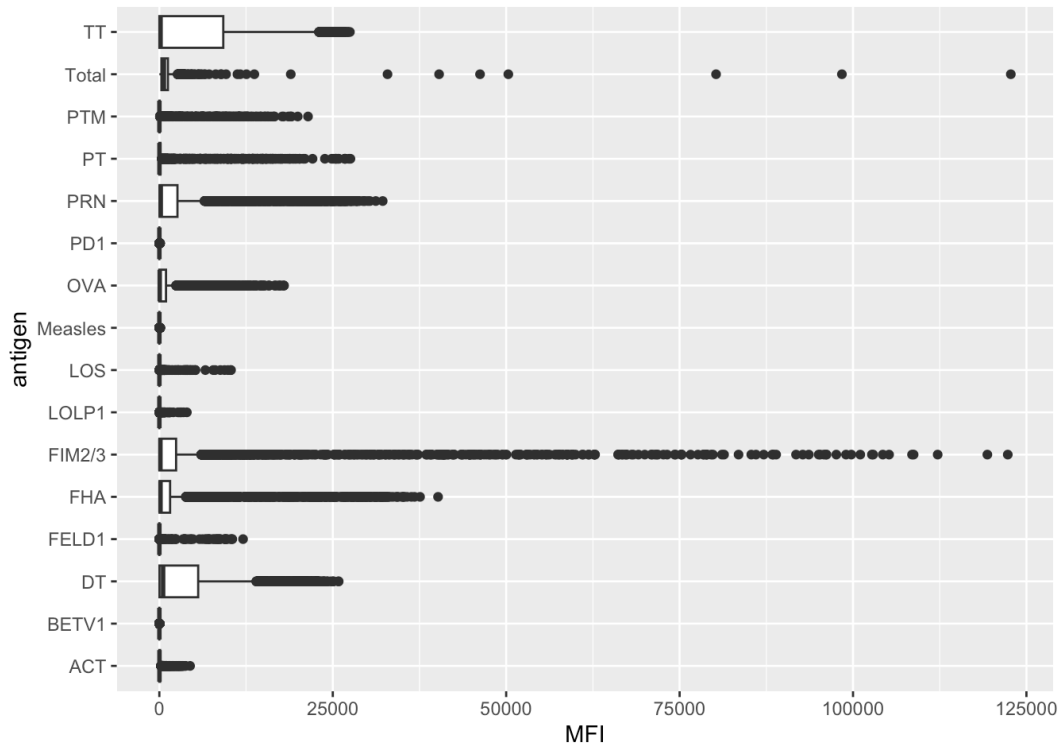
Q12

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

ACT	BETV1	DT	FELD1	FHA	FIM2/3	LOLP1
LOS Measles		OVA				
1970	1970	3435	1970	3829	3435	1970
1970	1970	3435				
PD1	PRN	PT	PTM	Total	TT	
1970	3829	3829	1970	788	3435	

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

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



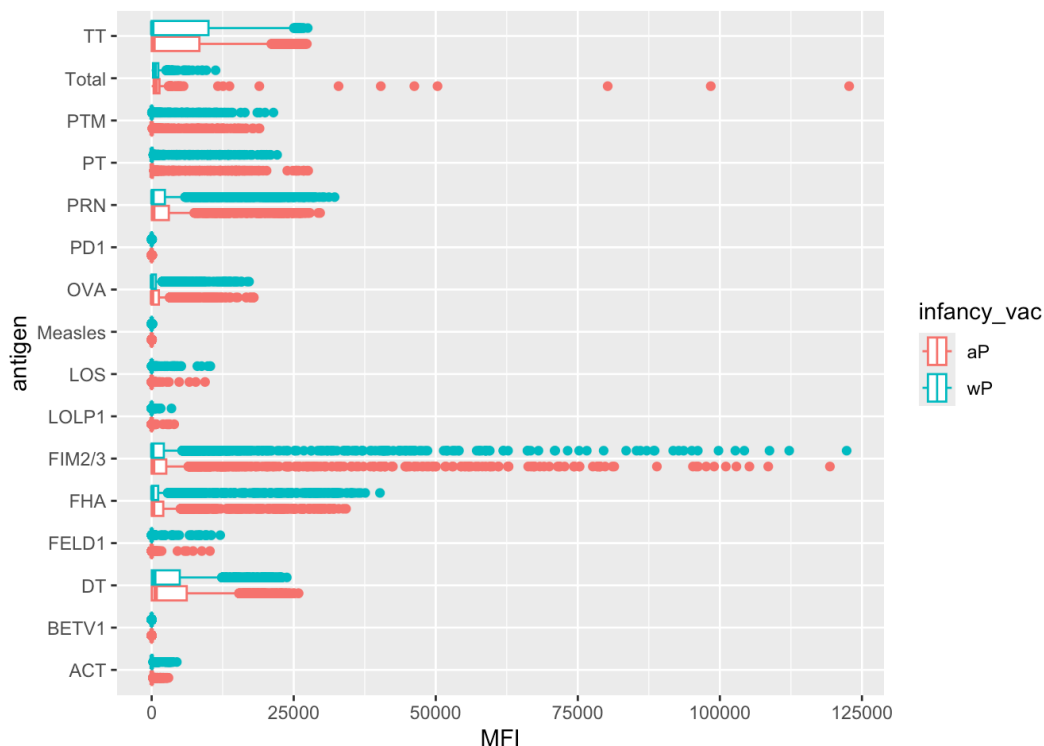
Q13. Complete the following code to make a summary boxplot of Ab titer levels (MFI) for all antigens:

why are certain antigens and not others very variable in their detection levels here?

can you facet or even just color by infancy_vac? Is there some difference?

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

Warning: Removed 1 row 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?

You can use the CMI-PB website search functionality and Terminology Browser (under development) to find out about each antigen. Note that this is still work in progress.

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
```

Lets focus in on just the 2021 dataset

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

2020_dataset 2021_dataset 2022_dataset
31520 8085 2170

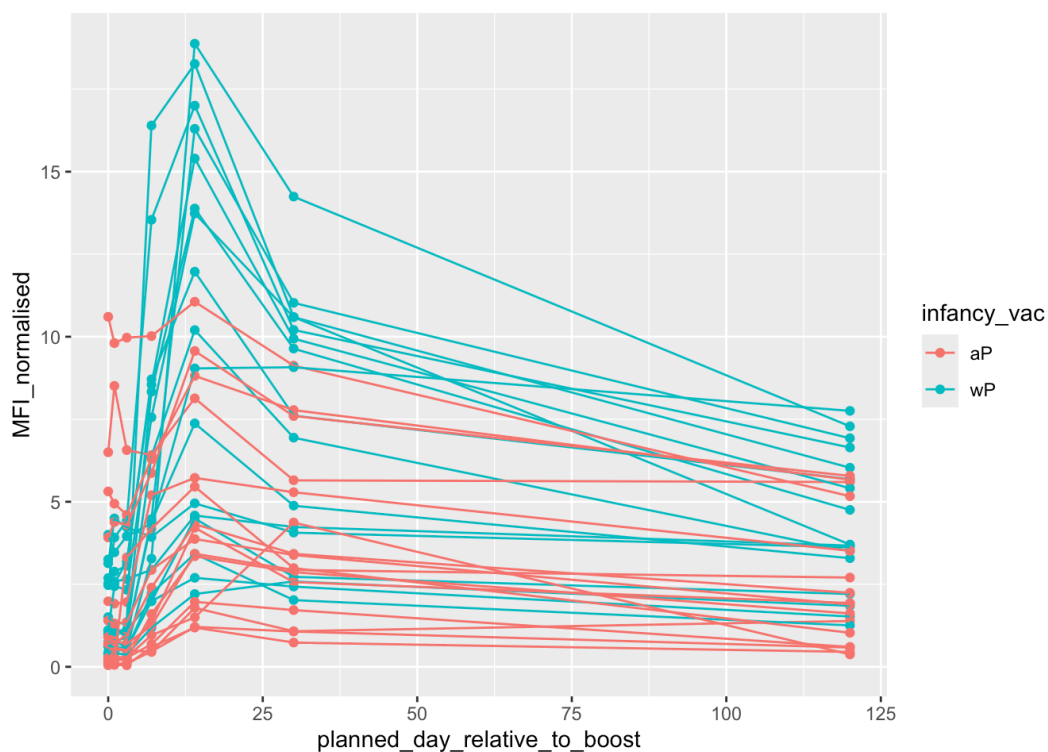
Focus on PT antigen IgG levels

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

plot of days(time) relative to boost vs MFI levels

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*).

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



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

PT levels clearly rise over time and far exceed those of OVA. They also appear to peak at visit 5 and then decline. This trend appears similar for wP and aP subjects.

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

Yes I do see differences in an increased MFI values in the beginning planned day to relative boost. ``