Lab 18- Pertussis mini-project

Emily Ignatoff (A16732102)

Pertussis (aka Whooping Cough) is a deadly lung infection caused by the bacteria *Bordatella pertussis*.

The CDC tracks pertussis cases around the US: https://tinyurl.com/pertussiscdc

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

```
head(cdc)
```

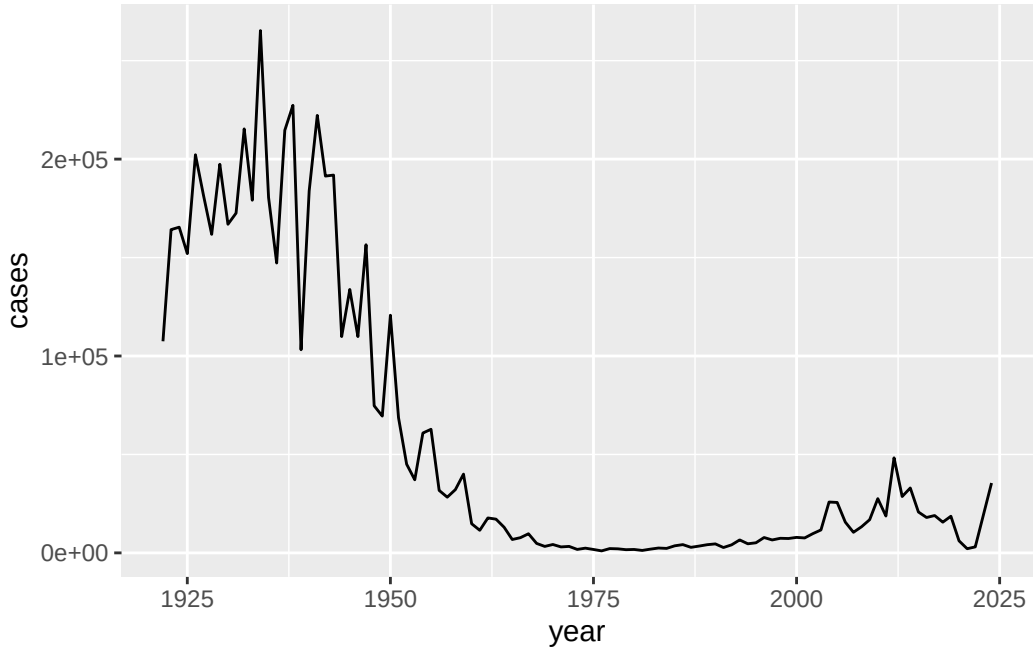
```
  year  cases
1 1922 107473
2 1923 164191
3 1924 165418
4 1925 152003
5 1926 202210
6 1927 181411
```

Q1 Let's make a ggplot with these data:

```
library(ggplot2)
```

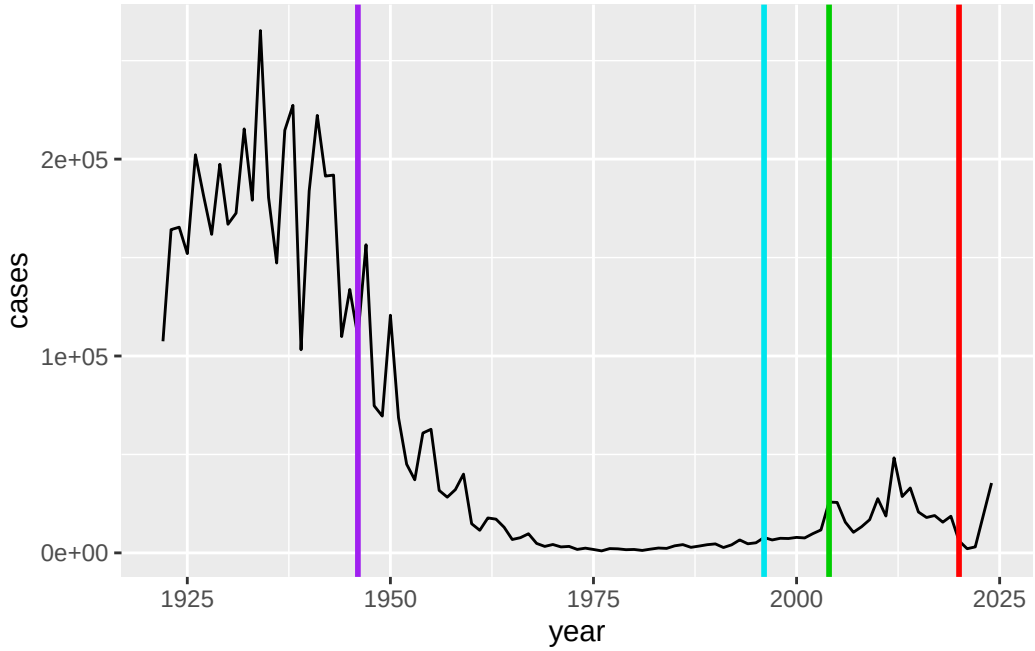
```
Warning: package 'ggplot2' was built under R version 4.3.3
```

```
ggplot(cdc) + aes(x=year, y=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?

```
ggplot(cdc) + aes(x=year, y=cases) +
  geom_line() + geom_vline(xintercept=1946, color="purple", lwd=1) +
  geom_vline(xintercept=1996, color= "turquoise2", lwd=1) +
  geom_vline(xintercept=2020, color= "red", lwd=1) +
  geom_vline(xintercept=2004, color="green3", lwd=1)
```



There were high case numbers before the first wP (whole cell) vaccine in 1946. Case number dropped drastically in the years following vaccine introduction to a low in the 1970s. After the introduction of the aP vaccine, around 2004, cases begin to peak again. Cases dip during the COVID pandemic, and rise rapidly following it.

> 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, cases start to trend upwards again (but nowhere near as high as before). Various factors can contribute to this, such as immunity waning without a booster shot, varying vaccination rates, and varying exposure to other people such as during quarantine. Perhaps the aP vaccine interacts with the immune system differently than the wP vaccine.

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

##CMI-PB (Computational Models of Immunity: Pertussis Boost)

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 project via JSON API calls To do this we will install the **jsonlite** package

```
library(jsonlite)
```

```
Warning: package 'jsonlite' was built under R version 4.3.3
```

```
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
```

Q How many individuals (subjects) are in this dataset?

```
nrow(subject)
```

```
[1] 172
```

Q. How many wP and aP primed individuals are in this dataset?

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

```
aP wP
87 85
```

Q. How many male and female subjects are in the study?

```
table(subject$biological_sex)
```

```
Female   Male
   112     60
```

```
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
```

This is not representative of the US population

Let's get the rest of the data we want:

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

```
  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 three tables of data from the CMI-PB project: `subject`, `specimen`, and `antibody`.

How can we put all these back together? We can use the **dplyr** functions `*_join()` In our case we will use an `inner_join()` for the data to avoid keeping individuals for whom follow-up testing was not possible.

```
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 specimen_id
1    1986-01-01    2016-09-12 2020_dataset           1
2    1986-01-01    2016-09-12 2020_dataset           2
3    1986-01-01    2016-09-12 2020_dataset           3
4    1986-01-01    2016-09-12 2020_dataset           4
5    1986-01-01    2016-09-12 2020_dataset           5
6    1986-01-01    2016-09-12 2020_dataset           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   8
```

```
dim(specimen)
```

```
[1] 1503    6
```

```
dim(meta)
```

```
[1] 1503   13
```

Now we can join the `antibody` data to the `meta` table we just generated:

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

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

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

```
  IgE   IgG  IgG1  IgG2  IgG3  IgG4
 6698  7265 11993 12000 12000 12000
```

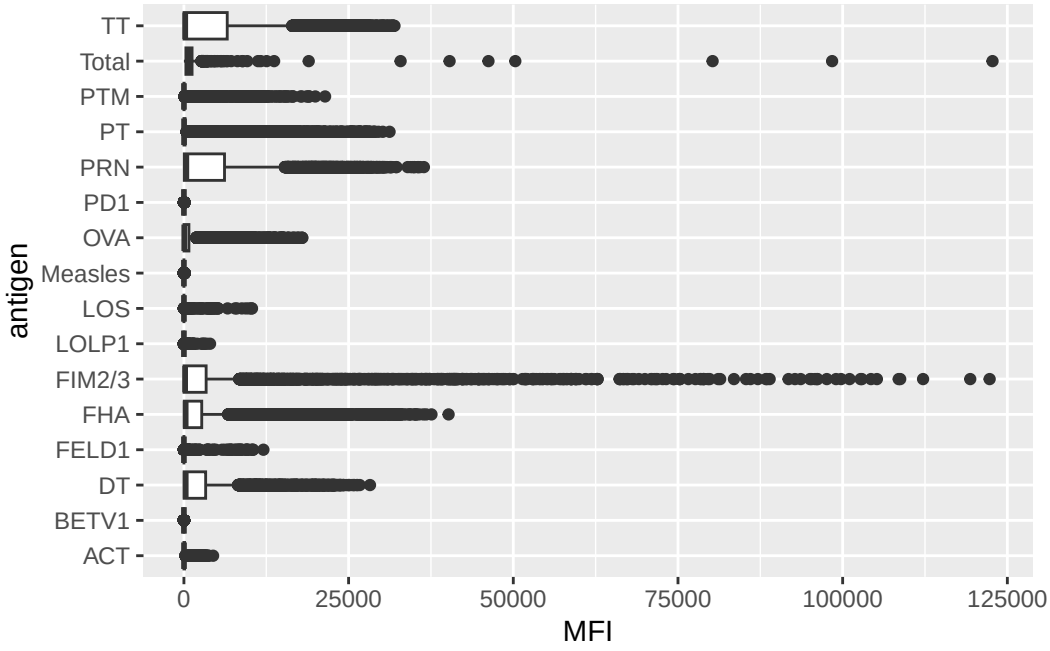
```
table(abdata$antigen)
```

```
    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
```

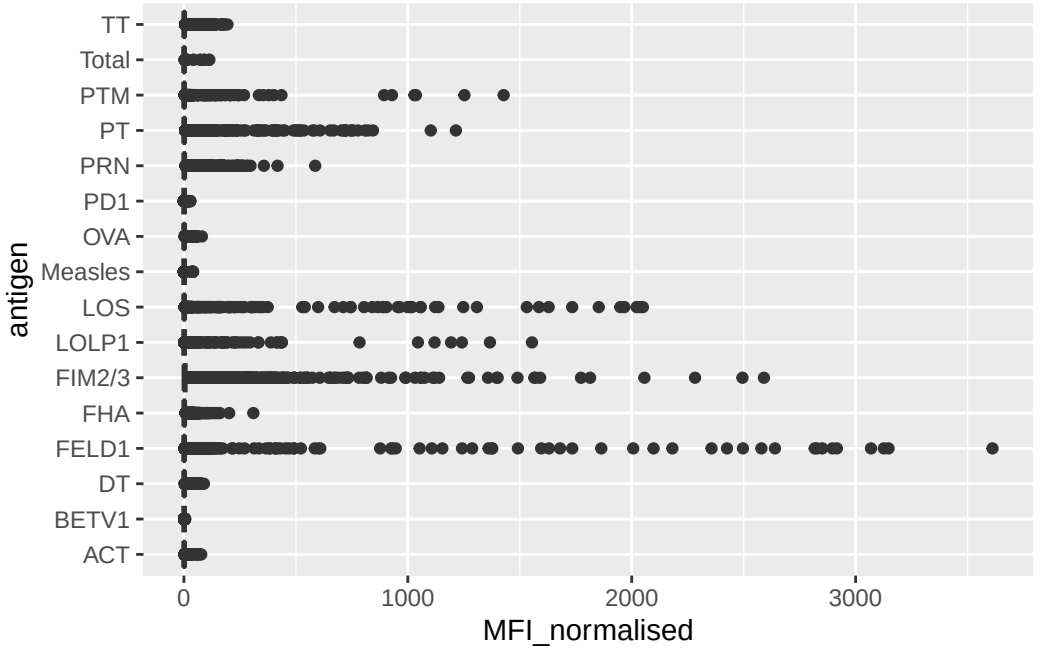
I want a plot of antigen levels across the whole dataset.

```
ggplot(abdata) + aes(x=MFI, y=antigen) +
  geom_boxplot()
```

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



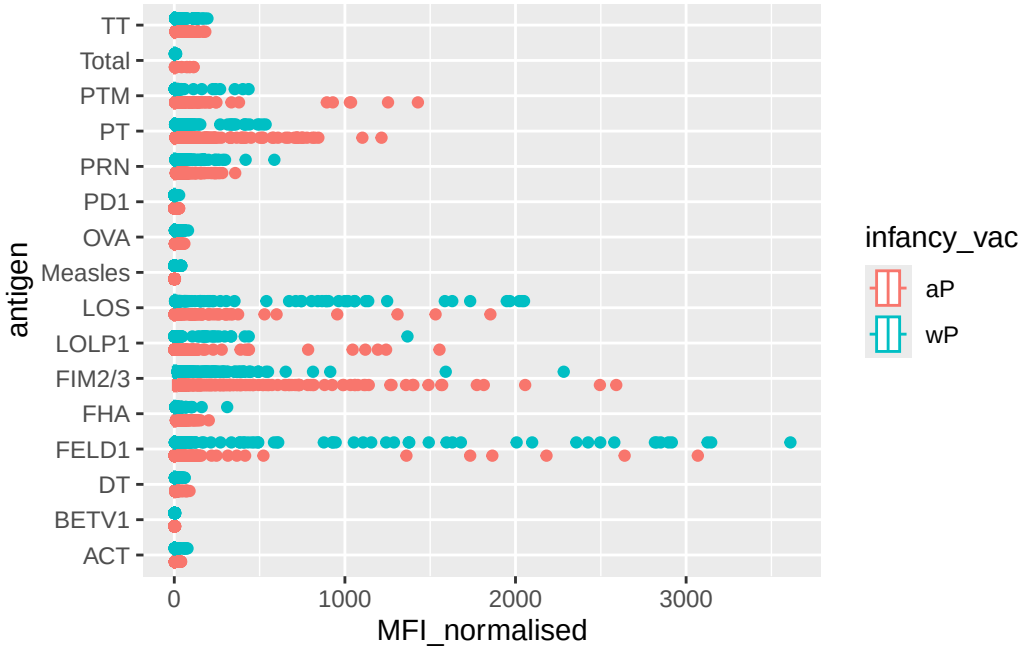

```
ggplot(abdata) + aes(x=MFI_normalised, y=antigen) +
  geom_boxplot()
```



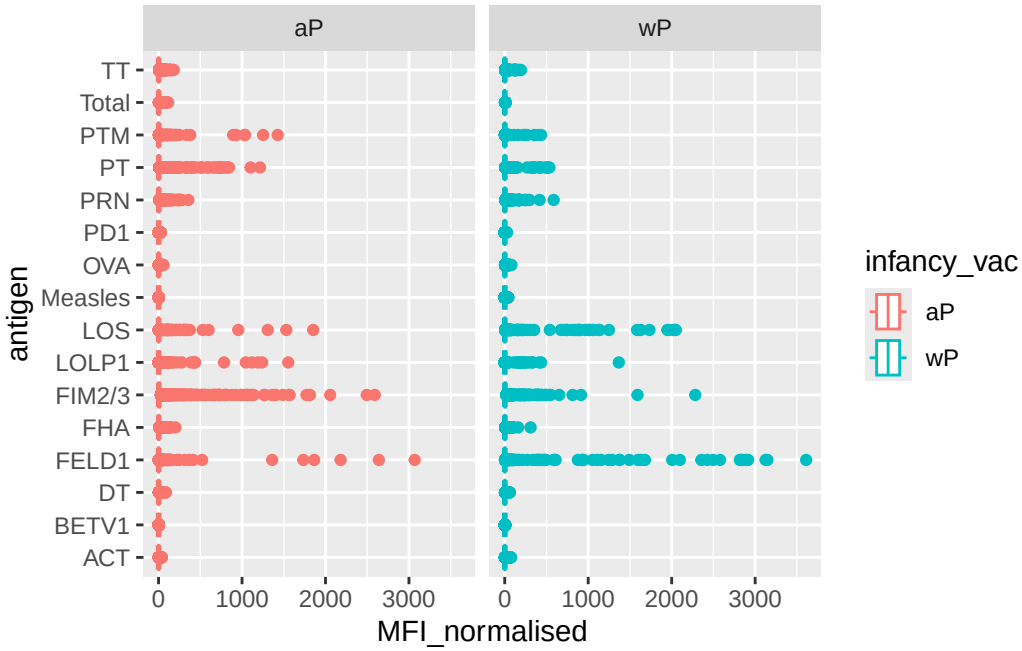
Antigens like FIM2/3, PT, FELD1 have quite a large range of values. Others like measles don't show much activity.

> Q. Are there differences at this whole-dataset level between aP and wP vaccinated individuals?

```
ggplot(abdata) + aes(x=MFI_normalised, y=antigen, col=infancy_vac) +
  geom_boxplot()
```




```
ggplot(abdata) + aes(x=MFI_normalised, y=antigen, col=infancy_vac) +
  geom_boxplot() + facet_wrap(~infancy_vac)
```



Examine IgG Ab titer levels

For this we need to isolate just isotype IgG.

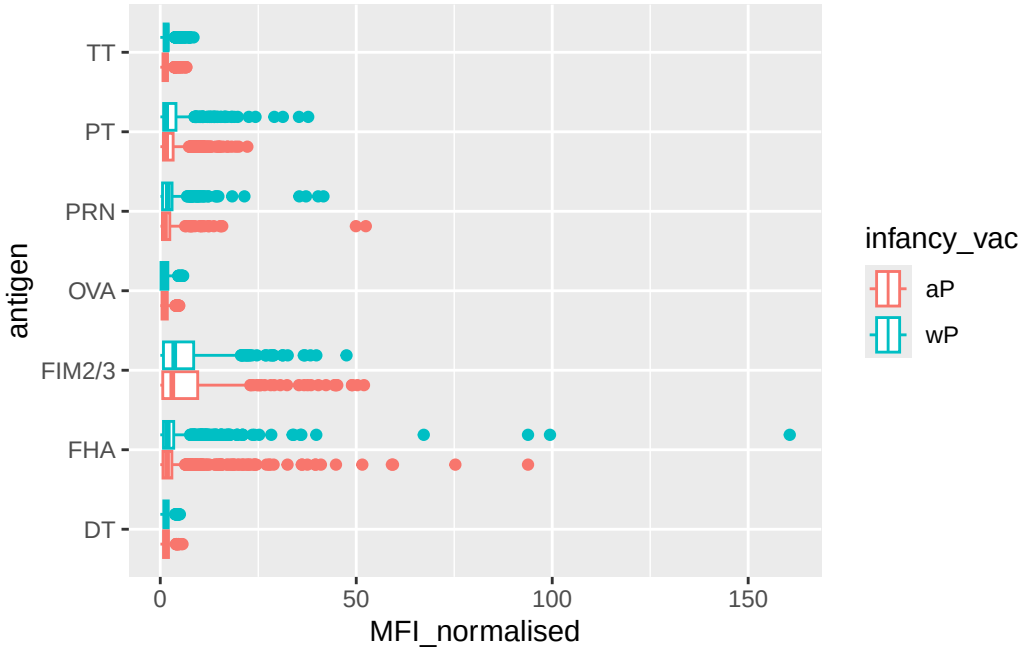
```
igg <- abdata %>% filter(isotype == "IgG")
head(igg)
```

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

An overview boxplot:

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

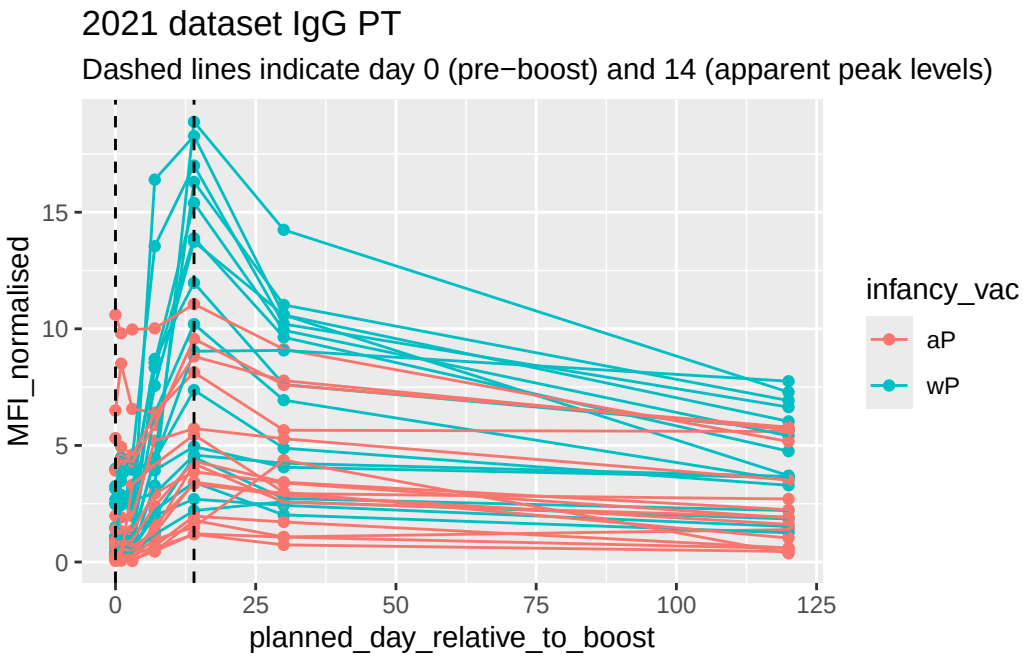


Digging further to look at the time course of IgG isotype PT antigen levels across aP and wP individuals

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

#filter to look at igg and pt only
pt.igg <- abdata.21 %>%
  filter(isotype == "IgG",  antigen == "PT")

#plot 2021 data, colored by infancy vac (aP vs wP)
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)")
```


2021 dataset IgG PT
Dashed lines indicate day 0 (pre−boost) and 14 (apparent peak levels)
MFI_normalised
15
10
5
0
0
25
50
75
100
125
planned_day_relative_to_boost
infancy_vac
aP
wP