Class18

Katie Mostoller A17259578

Table of contents

Pertussis (aka whooping cough) is a serious lung infection caused by the bacteria *bordetella pertussis*.

The CDC tracks Pertussis case numbers and we can find this data here: http://tinyurl.com/pertussiscdc

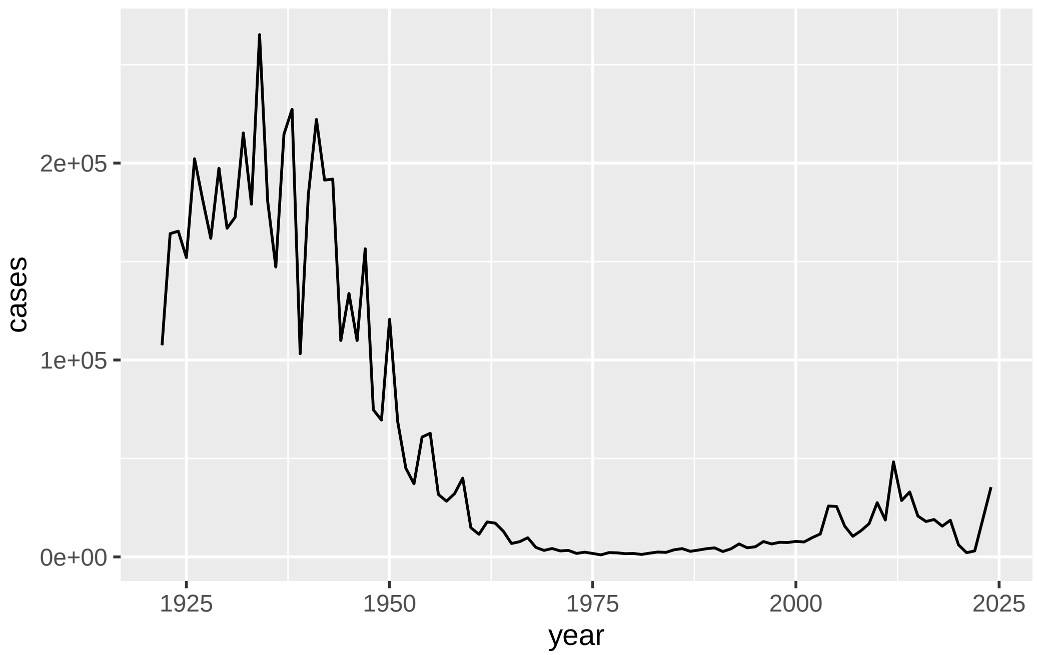
We can "scrape" this data using the **datapasta** package.

```r
head(cdc)
```

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

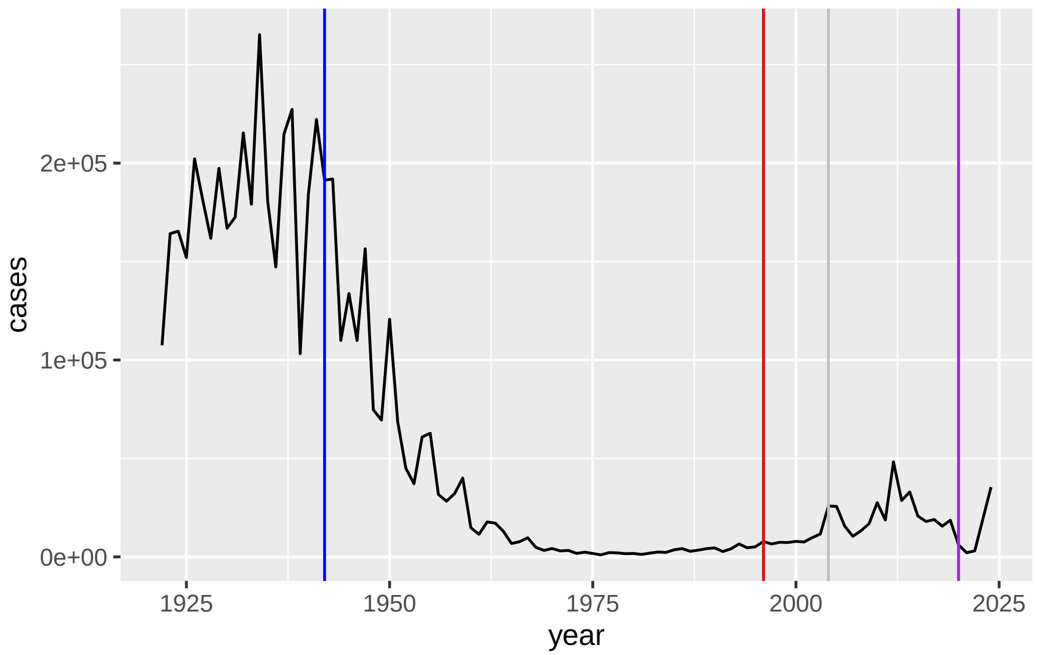
> Q1. Make a plot of pertussis cases per year using ggplot

```r
library(ggplot2)
ggplot(cdc) +
  aes(x = year, y = cases) +
  geom_line()
```



Q2. Let's add the key milestones of hte DTP (wP) vaccine roll out in 1942 and switch to the new aP vaccine in 1996. We can use `geom_vline()` for this. Booster shorts started in 2004.

```
ggplot(cdc) +
  aes(x = year, y = cases) +
  geom_line() +
  geom_vline(xintercept = 1942, col = "blue") +
  geom_vline(xintercept = 1996, col = "red") +
  geom_vline(xintercept = 2020, col = "purple") +
  geom_vline(xintercept = 2004, col = "grey")
```



There were high case numbers pre 1946 (before the sP vaccine) then relatively rapid decrease in case numbers through the 1970s to 2004 when our first widespread outbreak occurred again.

Noting the increase in yearly cases following the switch from to aP vaccine, there is suspiscion that aP vaccine induced immunity wanes faster than the older wP vaccine.

Enter the CMI-PB project

Computational Models of Immunity - Pertussis Boost

One of the main goals of this project is to determine what is different in the immune response between wP and aP primed individuals.

Using the booster vaccine as a proxy for infection

All data from this project is available here: https://www.cmi-pb.org/ in JSON format. We can use the **jsonlite()** package to read this data into R

```r
library(jsonlite)

subject <- read_json("http://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
```

> Q3. How many individuals are there in this dataset?

```r
nrow(subject)
```

```
[1] 172
```

> Q4. How many aP and wP individuals are there?

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

```

aP wP 
87 85 
```

> Q5. How many males and females are there?

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

```

Female   Male 
   112     60 
```

> Q6. Breakdown of biological sex and race?

```r
table(subject$rac, subject$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
```

> Q7. Does this look to be representative of the US population at large?

No, this information is largely pulled from UCSD student population

Let's read some more CMI-PB data

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

ab_titer <- read_json("http://cmi-pb.org/api/v5_1/plasma_ab_titer", simplifyVector = TRUE)
```

```r
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
```

```r
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
```

To use this data we need to "join" the various tables to find all the information we need to know about a particular measurement.

```
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
```

Now we can joing `meta` with `ab_titer`

```
ab_data <-  inner_join(meta, ab_titer)
```

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

```
head(ab_data)
```

```
  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
```

Q8. How many different antibody isotypes are we measuring?

```
table(ab_data$isotype)
```

```

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

Q8. How many different antigens are we measuring?

```
table(ab_data$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 
```

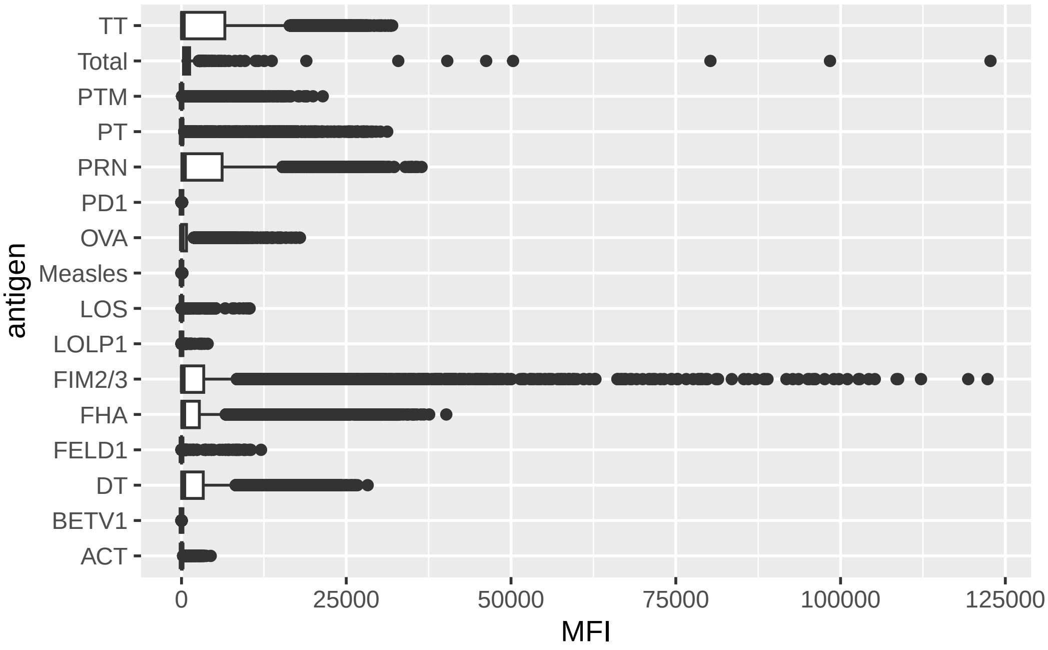
Q9. Let's look at a boxplot of antigen levels over the whole dataset?

```
dim(ab_data)
```

```
[1] 61956    20
```

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

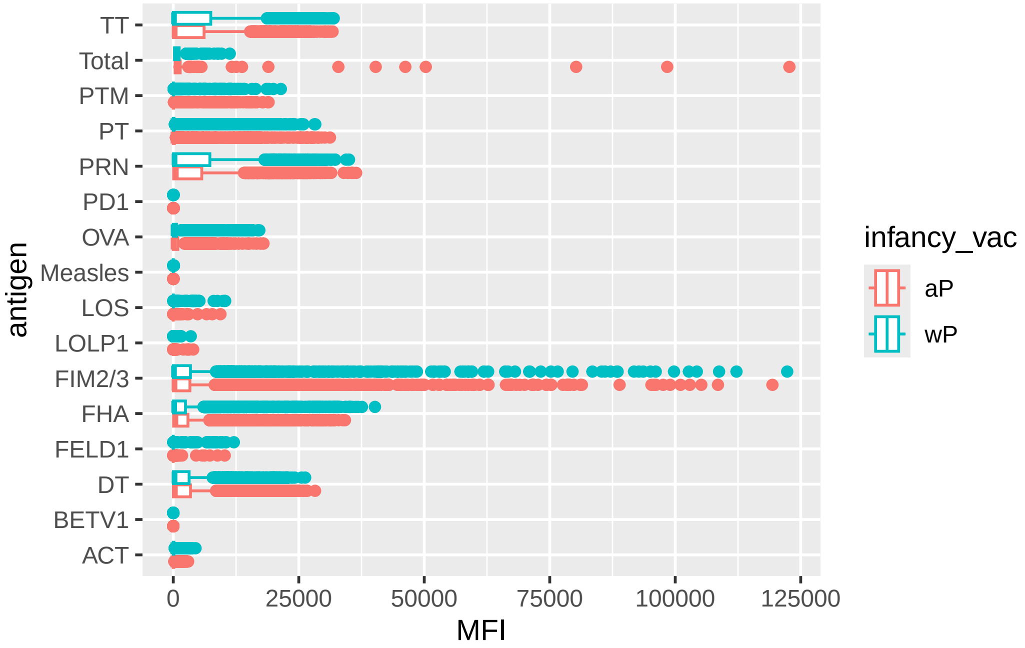
```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_boxplot()`).
```



Q10. Break this plot down by aP or wP

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

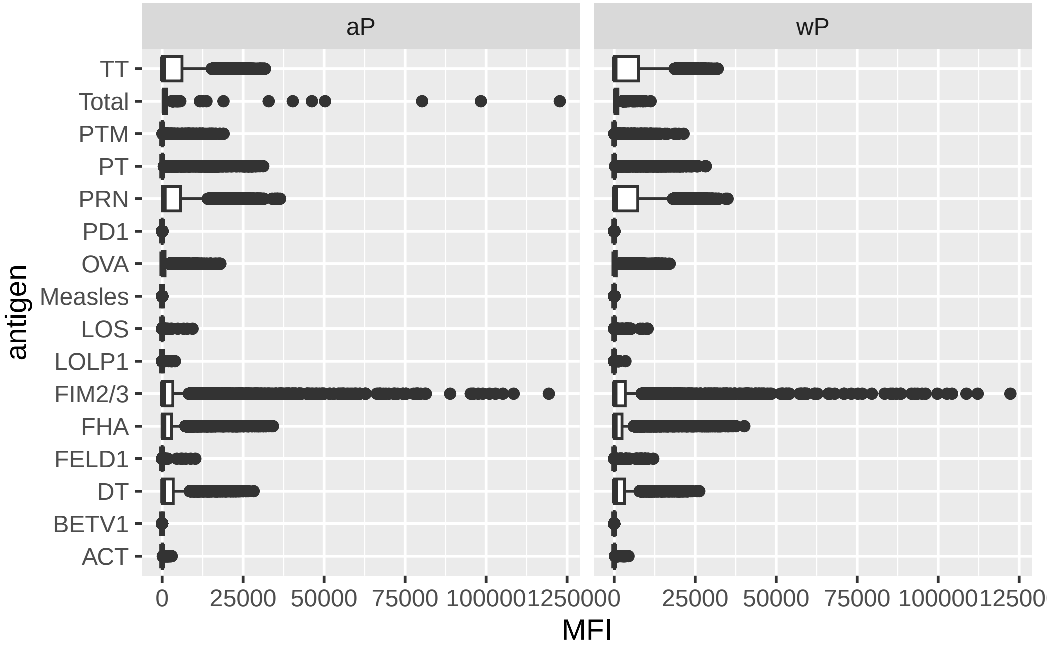
```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_boxplot()`).
```



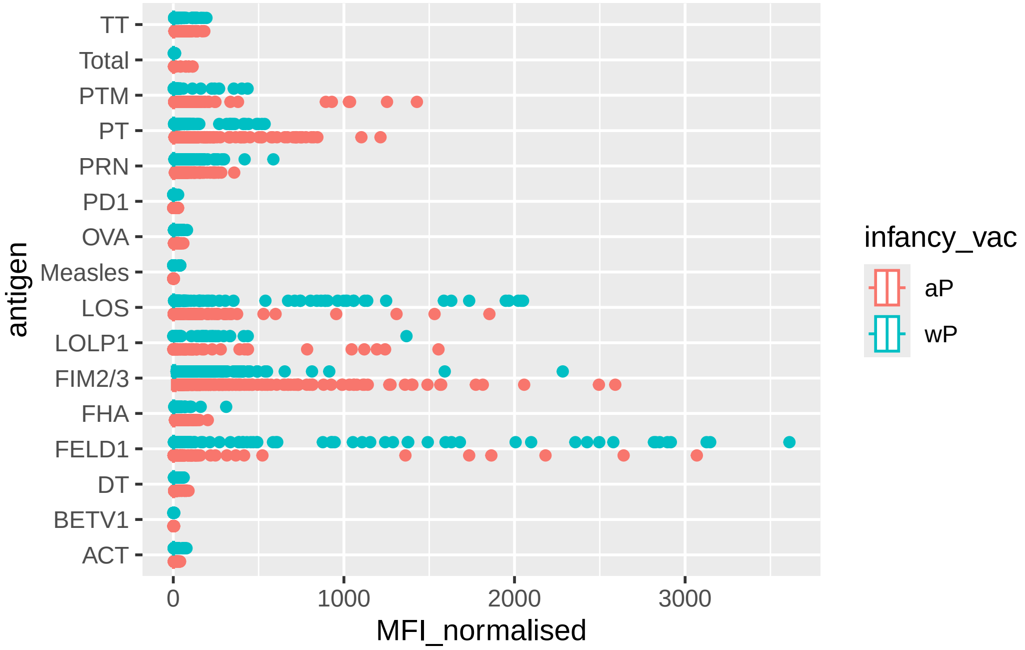
We can facet the plot by `infancy_vac`

```
ggplot(ab_data) +
  aes(MFI, antigen) +
  geom_boxplot() +
  facet_wrap(~infancy_vac)
```

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



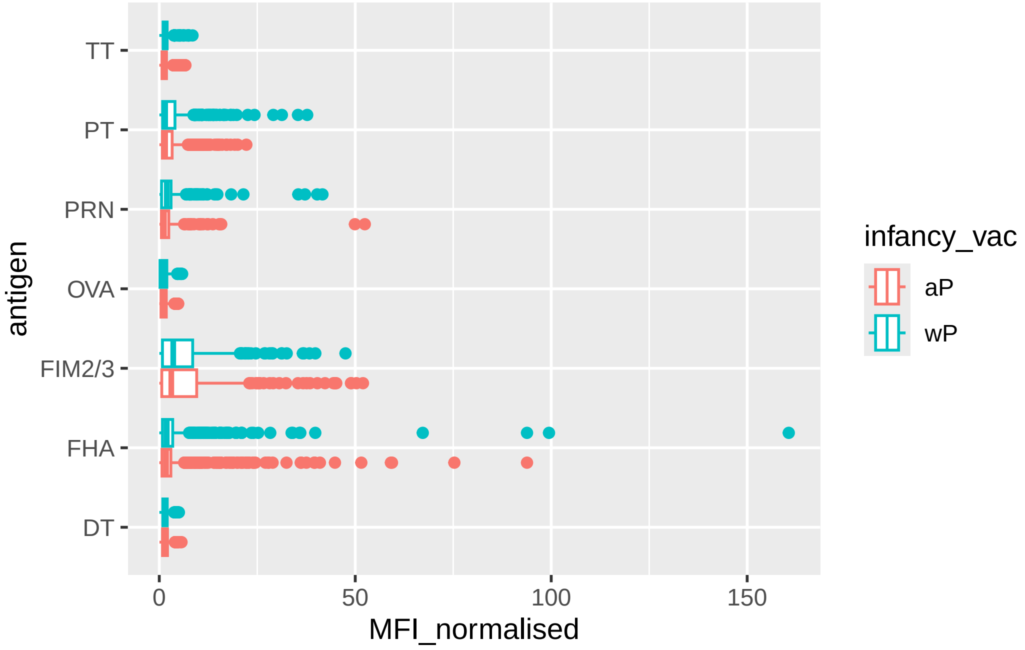
```
ggplot(ab_data) +
  aes(MFI_normalised, antigen, col = infancy_vac) +
  geom_boxplot()
```



Let's focus on just IgG

```
igg <- ab_data |>
       filter(isotype=="IgG")
```

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

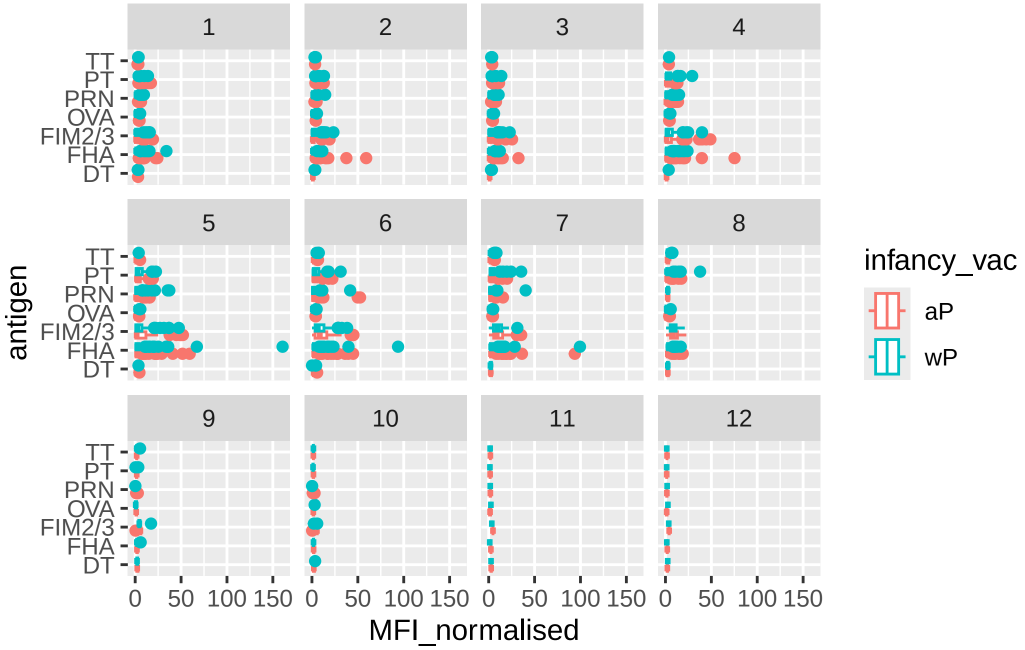


```
head(igg)
```

```
  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
```

```r
ggplot(igg) +
  aes(MFI_normalised, antigen, col = infancy_vac) +
  geom_boxplot() +
  facet_wrap(~visit)
```



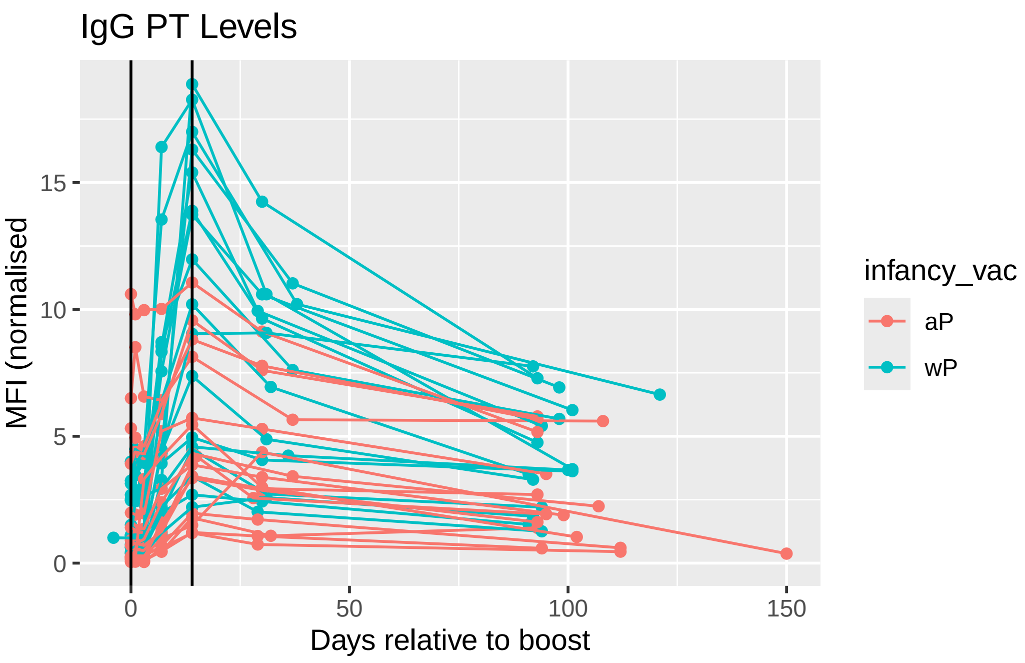
Let's focus on PT (pertussis toxin) and IgG over time

Filter to focus on one antigen (PT) and IgG levels for one of the datasets

```
pt_igg <- ab_data |>
        filter(isotype=="IgG", antigen=="PT", dataset=="2021_dataset")
```

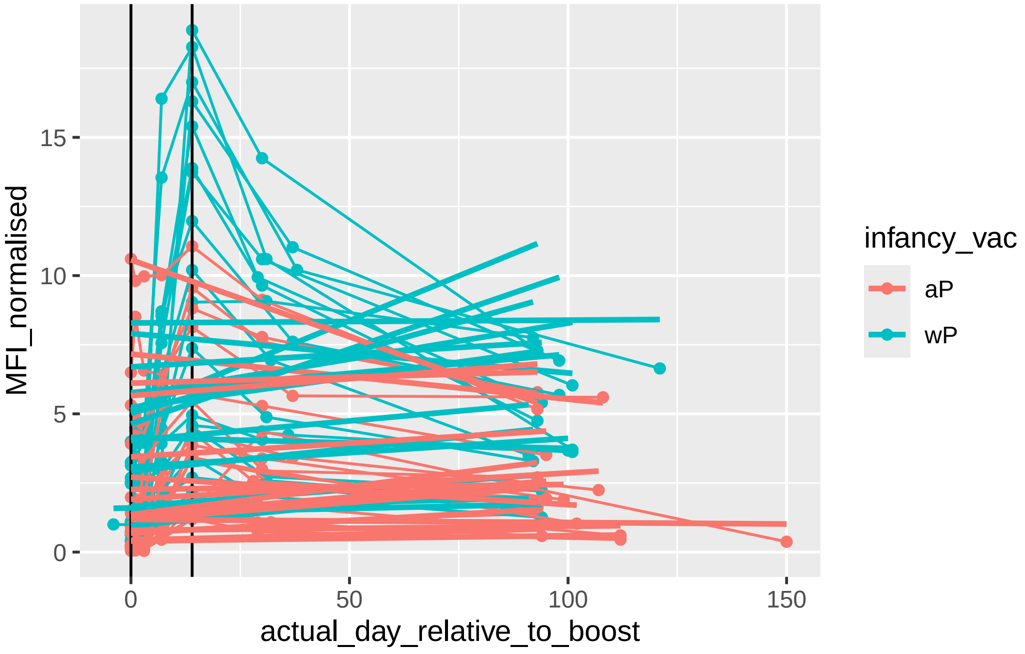
A plot of `actual_day_relative_to_boost` vs `MFI_normalised`

```
ggplot(pt_igg) +
  aes(actual_day_relative_to_boost, MFI_normalised, col = infancy_vac, group = subject_id) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 14) +
  geom_vline(xintercept = 0) +
  labs(title="IgG PT Levels", x="Days relative to boost", y="MFI (normalised")
```



```
ggplot(pt_igg) +
  aes(actual_day_relative_to_boost, MFI_normalised, col = infancy_vac, group = subject_id) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 14) +
  geom_vline(xintercept = 0) +
  geom_smooth(method = "lm", se = FALSE)
```

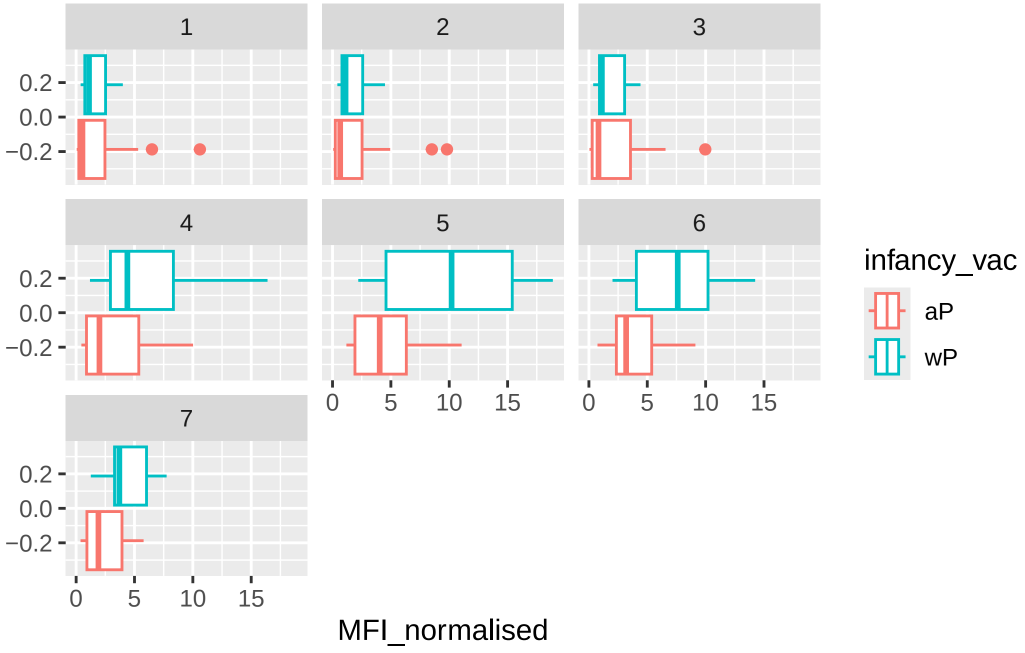
```
`geom_smooth()` using formula = 'y ~ x'
```



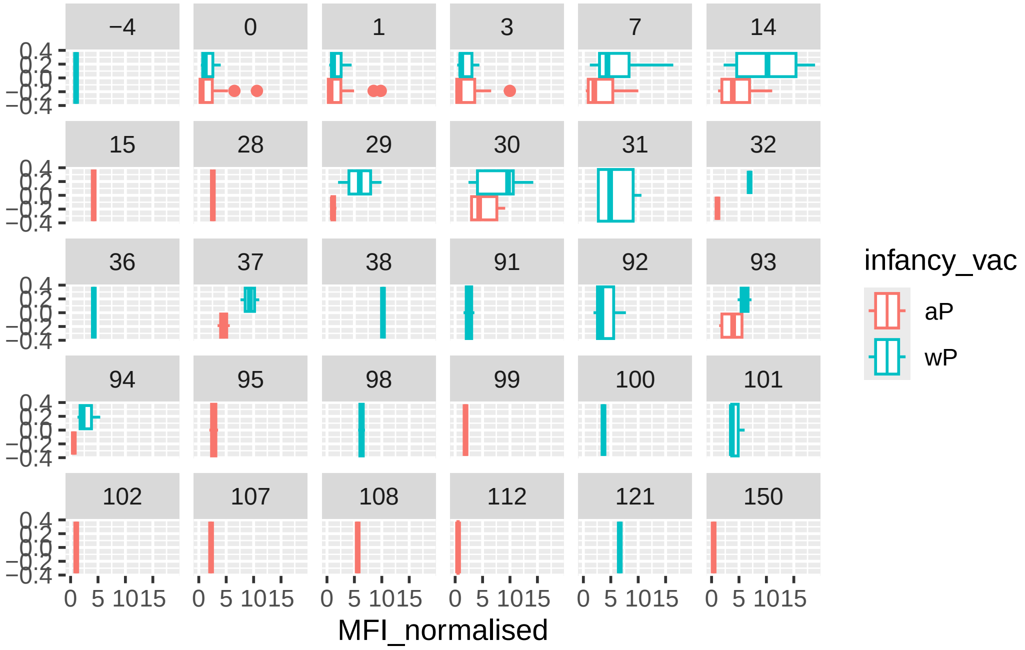
```
  labs(title="IgG PT Levels", x="Days relative to boost", y="MFI (normalised")
```

```
$x
[1] "Days relative to boost"

$y
[1] "MFI (normalised"

$title
[1] "IgG PT Levels"

attr(,"class")
[1] "labels"
```

```
ggplot(pt_igg) +
  aes(MFI_normalised, col = infancy_vac) +
  geom_boxplot() +
  facet_wrap(~visit)
```



```
ggplot(pt_igg) +
  aes(MFI_normalised, col = infancy_vac) +
  geom_boxplot() +
  facet_wrap(~actual_day_relative_to_boost)
```



```r
library(lubridate)
```

```

Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

```r
today() - mdy("12-12-1997")
```

```
Time difference of 9947 days
```

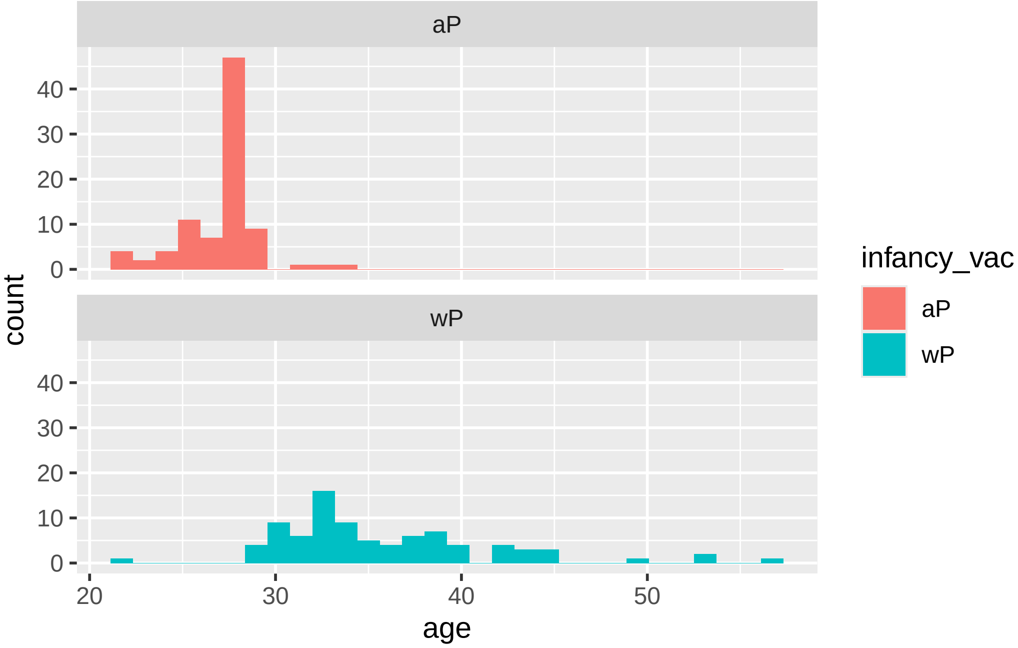
```r
time_length(today() - mdy("12-12-1997"), "years")
```

```
[1] 27.2334
```

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

```
ggplot(subject) +
  aes(age, fill = infancy_vac) +
  geom_histogram() +
  facet_wrap(~infancy_vac, ncol = 1)
```

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


aP
wP
count
age
infancy_vac
aP
wP