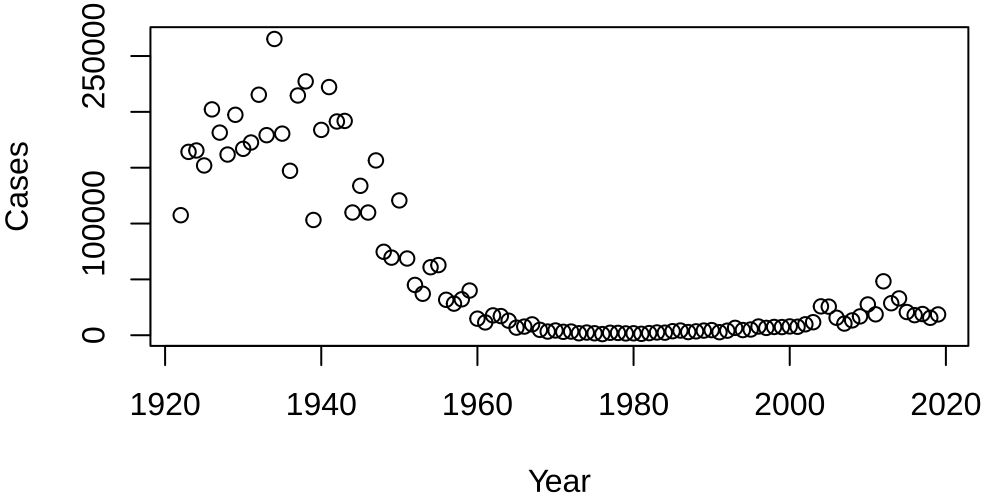
Class 19:Vaccine Mini-Project

Suzanne Enos

Web scraping

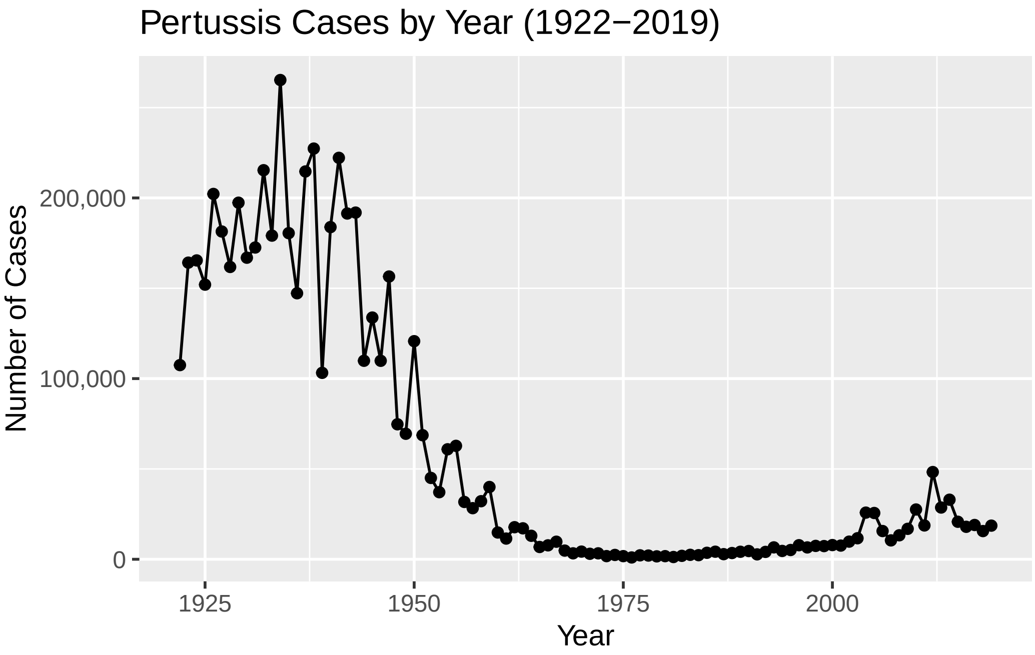
Extracting the CDC figures for Pertussis cases in the US: https://www.cdc.gov/pertussis/surv-reporting/cases-by-year.html Using datapasta to paste data from website as a datafram.

```
plot(cdc)
```



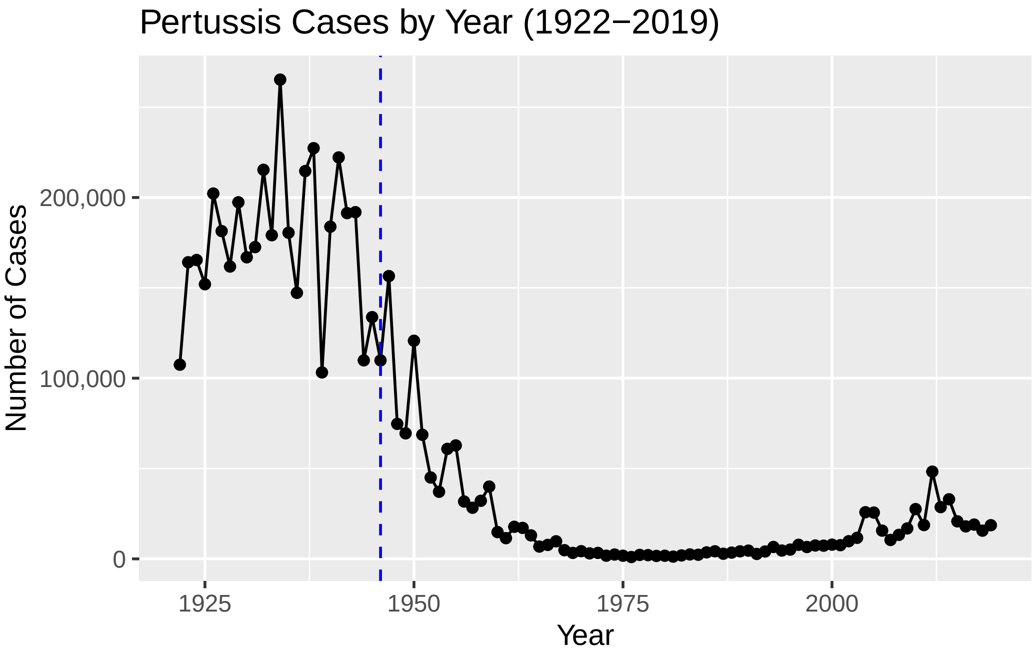
```
library(ggplot2)

base <- ggplot(cdc) +
  aes(Year, Cases) +
  geom_point() +
  geom_line() +
  labs(title = "Pertussis Cases by Year (1922-2019)", y = "Number of Cases") +
  scale_y_continuous(labels = scales::label_comma())
base
```



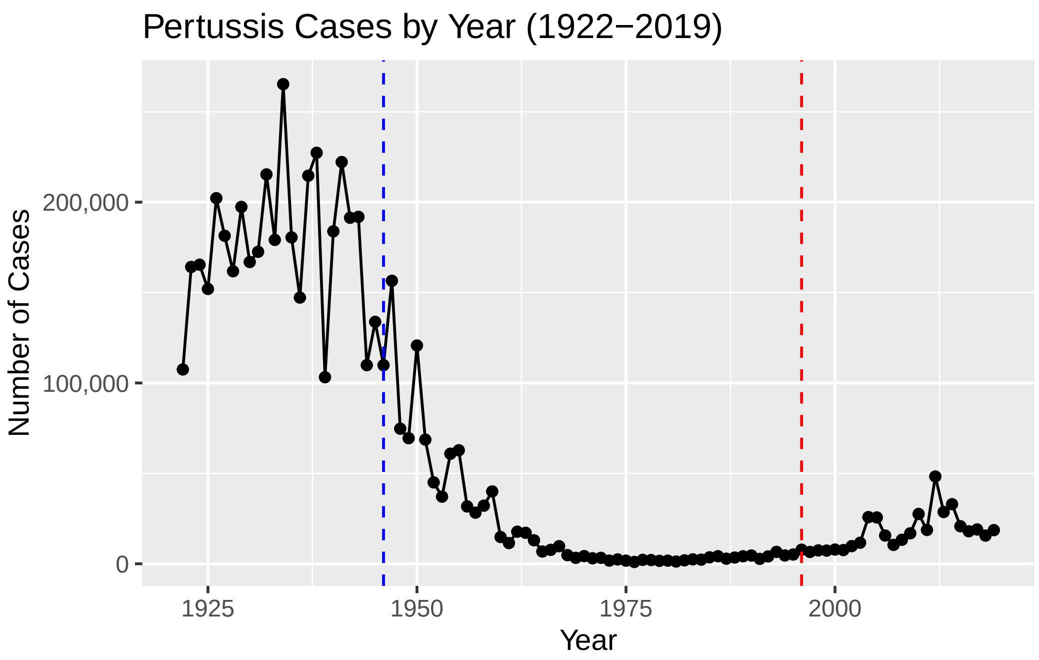
First vaccine in 1946

```
base +
  geom_vline(xintercept = 1946, col = "blue", linetype = 2)
```



New vaccine formulation in 1996 - wP to aP (whole-cell to acellular) US and many other countries switched. Not all countries switched.

```
base +
  geom_vline(xintercept = 1946, col = "blue", linetype = 2) +
  geom_vline(xintercept = 1996, col = "red", linetype = 2)
```



Exploring CMI-PB data

Why is this vaccine-preventable disease on the upswing?

CMI-PB resource API returns JSON format. We will use **jsonlite**

```
library(jsonlite)

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

How many wP and aP subjects are there?

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

```

aP wP 
47 49 
```

How many female non-white individuals are there in the dataset?

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

```
                                           
                                            Female Male
  American Indian/Alaska Native                  0    1
  Asian                                         18    9
  Black or African American                      2    0
  More Than One Race                             8    2
  Native Hawaiian or Other Pacific Islander      1    1
  Unknown or Not Reported                       10    4
  White                                         27   13
```

Let's look at the specimen table

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

```
  specimen_id subject_id actual_day_relative_to_boost
1           1          1                           -3
2           2          1                          736
3           3          1                            1
4           4          1                            3
5           5          1                            7
6           6          1                           11
  planned_day_relative_to_boost specimen_type visit
```

```
1                              0         Blood      1
2                            736         Blood     10
3                              1         Blood      2
4                              3         Blood      3
5                              7         Blood      4
6                             14         Blood      5
```

```
dim(specimen)
```

```
[1] 729   6
```

```
dim(subject)
```

```
[1] 96  8
```

```
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(specimen, subject)
```

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

```
dim(meta)
```

```
[1] 729  13
```

```r
titer <- read_json("http://cmi-pb.org/api/ab_titer", simplifyVector = T)
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
```

```r
dim(titer)
```

```
[1] 32675     8
```

> How many different isotypes are there?

```r
table(titer$isotype)
```

```

 IgE  IgG IgG1 IgG2 IgG3 IgG4 
6698 1413 6141 6141 6141 6141 
```

Merge titer and meta

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

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

```r
dim(abdata)
```

```
[1] 32675    20
```

```r
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 actual_day_relative_to_boost
1 UG/ML                 2.096133          1                           -3
2 IU/ML                29.170000          1                           -3
3 IU/ML                 0.530000          1                           -3
4 IU/ML                 6.205949          1                           -3
5 IU/ML                 4.679535          1                           -3
6 IU/ML                 2.816431          1                           -3
  planned_day_relative_to_boost specimen_type visit infancy_vac biological_sex
1                             0         Blood     1          wP         Female
2                             0         Blood     1          wP         Female
3                             0         Blood     1          wP         Female
4                             0         Blood     1          wP         Female
5                             0         Blood     1          wP         Female
6                             0         Blood     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
```

> Q. What do you notice about “visit” number 8?

```r
table(abdata$visit)
```

```

   1    2    3    4    5    6    7    8 
5795 4640 4640 4640 4640 4320 3920   80 
```

Visit 8 still ongoing, still collecting data. Will just analyze visits 1-7.

Examine IgG1 Ab titer levels

```
ig1 <- filter(abdata, isotype == "IgG1", visit!= 8)
dim(ig1)
```

```
[1] 6126   20
```

```
head(ig1)
```

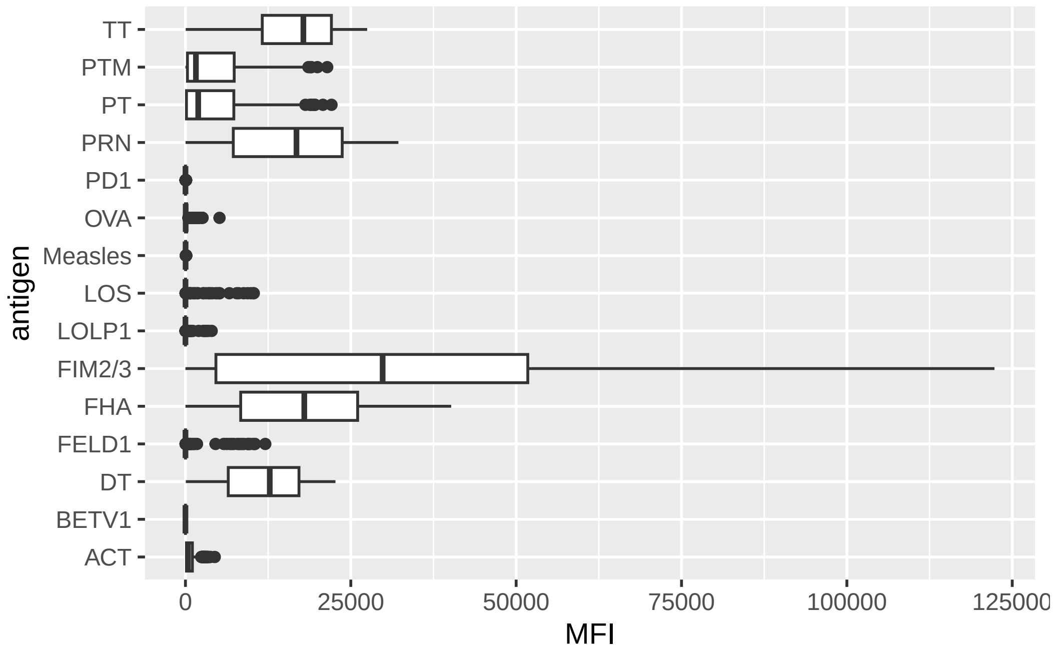
```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1    IgG1                TRUE     ACT 274.355068      0.6928058
2           1    IgG1                TRUE     LOS  10.974026      2.1645083
3           1    IgG1                TRUE   FELD1   1.448796      0.8080941
4           1    IgG1                TRUE   BETV1   0.100000      1.0000000
5           1    IgG1                TRUE   LOLP1   0.100000      1.0000000
6           1    IgG1                TRUE Measles  36.277417      1.6638332
   unit lower_limit_of_detection subject_id actual_day_relative_to_boost
1 IU/ML                 3.848750          1                           -3
2 IU/ML                 4.357917          1                           -3
3 IU/ML                 2.699944          1                           -3
4 IU/ML                 1.734784          1                           -3
5 IU/ML                 2.550606          1                           -3
6 IU/ML                 4.438966          1                           -3
  planned_day_relative_to_boost specimen_type visit infancy_vac biological_sex
1                             0         Blood     1          wP         Female
2                             0         Blood     1          wP         Female
3                             0         Blood     1          wP         Female
4                             0         Blood     1          wP         Female
5                             0         Blood     1          wP         Female
6                             0         Blood     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
```

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

```
    ACT   BETV1      DT   FELD1     FHA  FIM2/3   LOLP1     LOS Measles     OVA
   1970    1970    2135    1970    2529    2135    1970    1970    1970    2135
    PD1     PRN      PT     PTM   Total      TT
   1970    2529    2529    1970     788    2135
```

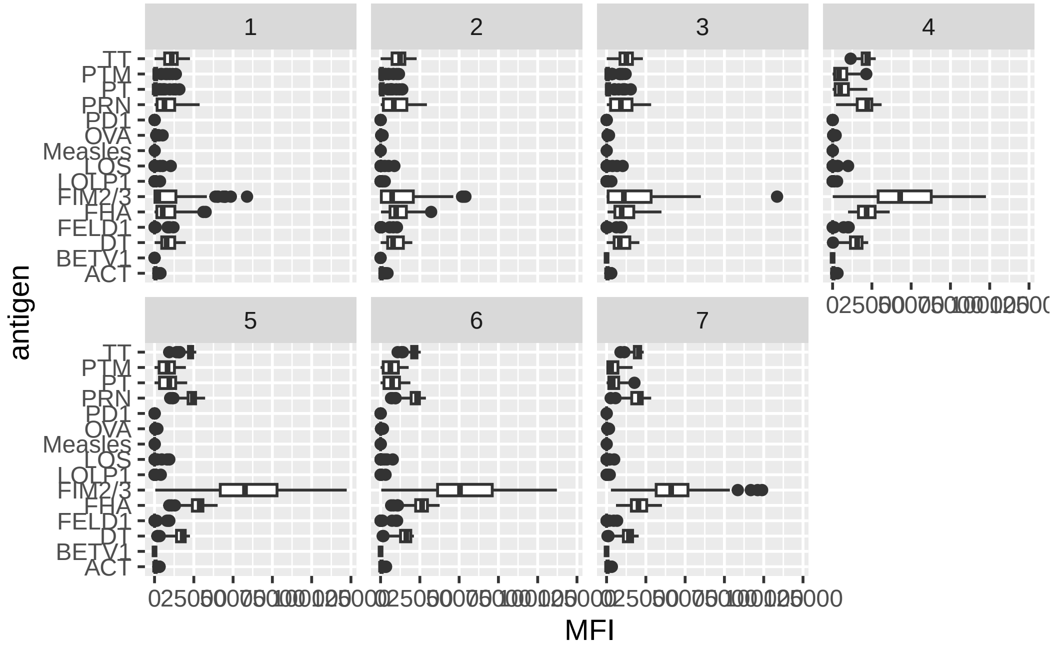
Analysis of the whole dataset: antigen levels (plot of antigen vs MFI)

```
ggplot(ig1) +
  aes(MFI, antigen) +
  geom_boxplot()
```



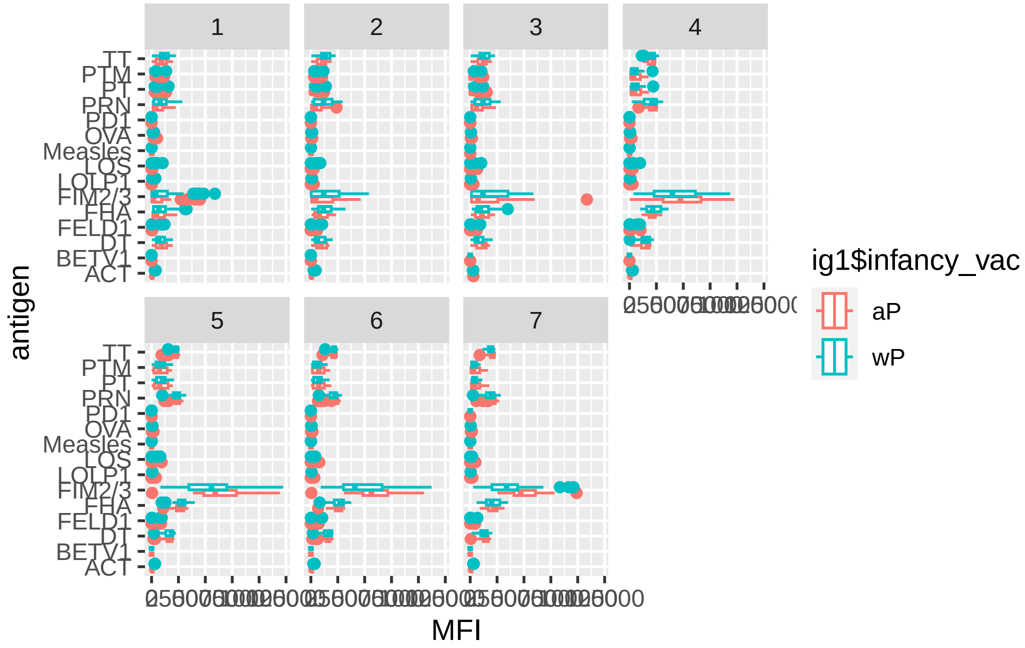
Add faceting by visit

```
ggplot(ig1) +
  aes(MFI, antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow = 2)
```

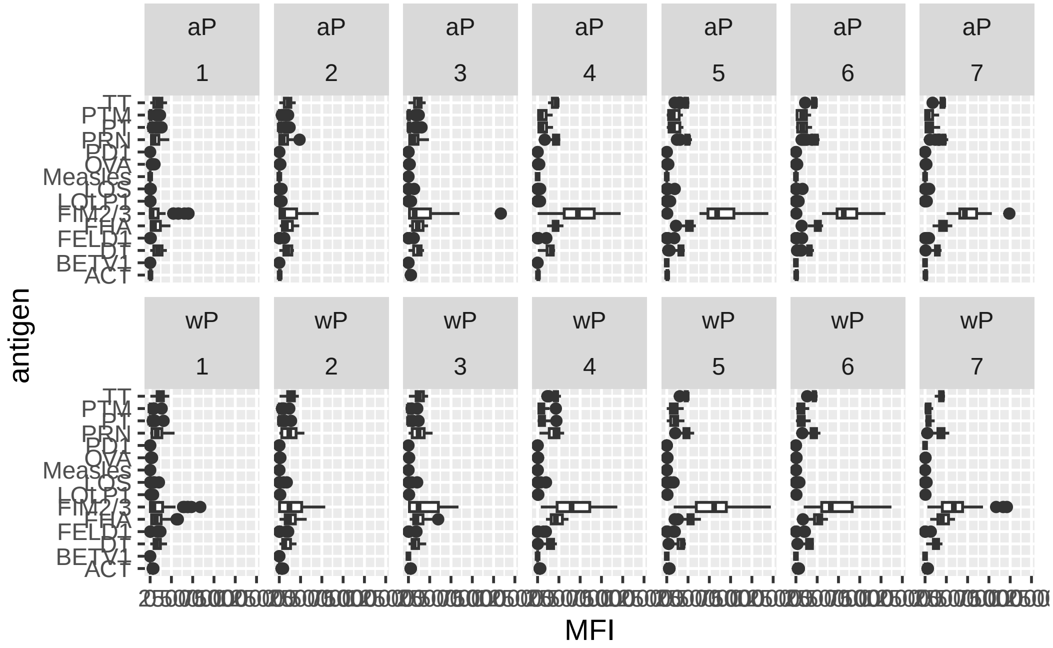


```
ggplot(ig1) +
  aes(MFI, antigen, col = ig1$infancy_vac) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow = 2)
```

```
Warning: Use of `ig1$infancy_vac` is discouraged.
i Use `infancy_vac` instead.
```

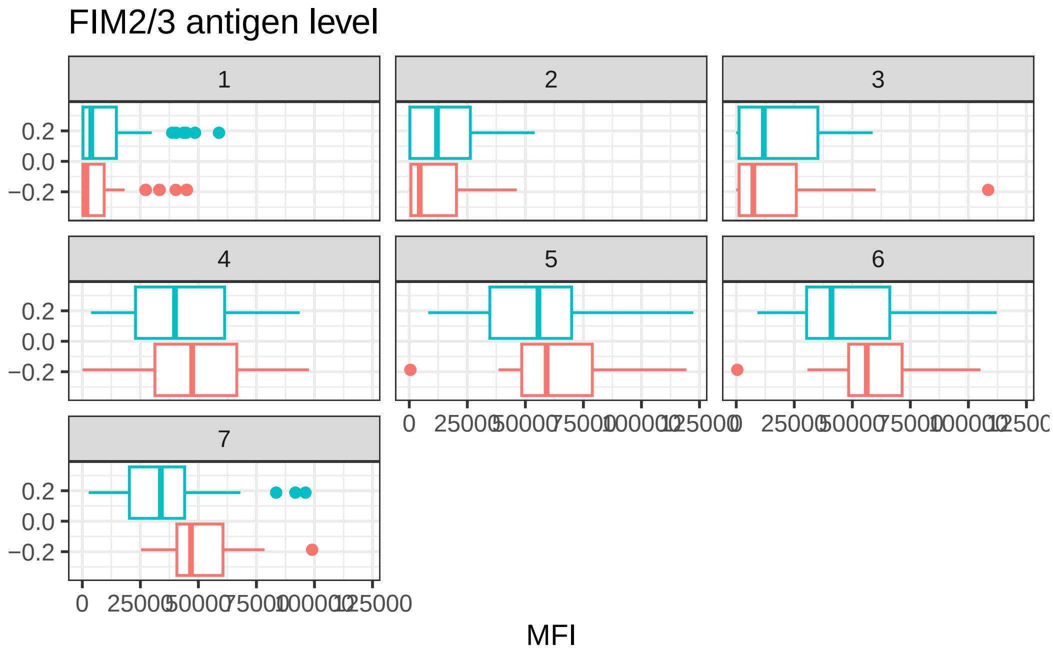


```
ggplot(ig1) +
  aes(MFI, antigen) +
  geom_boxplot() +
  facet_wrap(vars(infancy_vac, visit), nrow = 2)
```

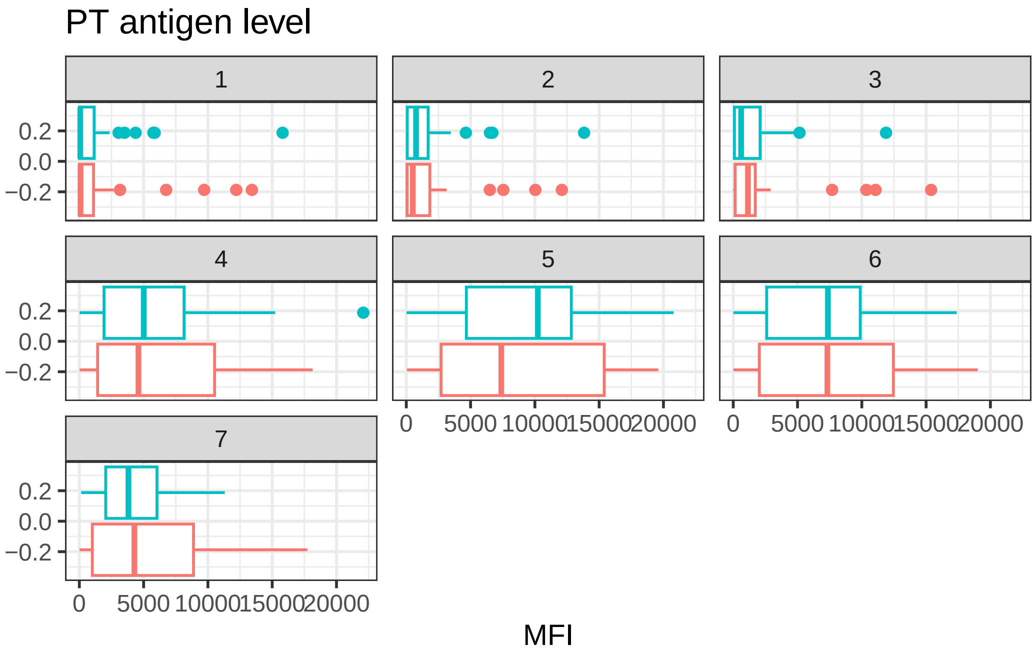


Unclear what the difference in aP and wP

```
filter(ig1, antigen== "FIM2/3") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  labs(title = "FIM2/3 antigen level") +
  theme_bw()
```



```
filter(ig1, antigen== "PT") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  labs(title = "PT antigen level") +
  theme_bw()
```



Do you see any clear differences in wP and aP response?

Not really

For RNA-Seq data the API query mechanism quickly hits the web browser interface limit for file size. We will present alternative download mechanisms for larger CMI-PB datasets in the next section. However, we can still do "targeted" RNA-Seq querys via the web accessible API.

For example we can obtain RNA-Seq results for a specific ENSEMBLE gene identifier or multiple identifiers combined with the & character:

```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.

rna <- read_json(url, simplifyVector = TRUE)
head(rna)
```

```
  versioned_ensembl_gene_id specimen_id raw_count      tpm
1         ENSG00000211896.7         344     18613  929.640
2         ENSG00000211896.7         243      2011  112.584
3         ENSG00000211896.7         261      2161  124.759
4         ENSG00000211896.7         282      2428  138.292
5         ENSG00000211896.7         345     51963 2946.136
6         ENSG00000211896.7         244     49652 2356.749
```

Join rna to meta because different specimen from Ab titer

```
ssrna <- inner_join(rna, meta)
```

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

```
head(ssrna)
```

```
  versioned_ensembl_gene_id specimen_id raw_count      tpm subject_id
1         ENSG00000211896.7         344     18613  929.640         44
2         ENSG00000211896.7         243      2011  112.584         31
3         ENSG00000211896.7         261      2161  124.759         33
4         ENSG00000211896.7         282      2428  138.292         36
5         ENSG00000211896.7         345     51963 2946.136         44
6         ENSG00000211896.7         244     49652 2356.749         31
  actual_day_relative_to_boost planned_day_relative_to_boost specimen_type
1                            3                             3         Blood
2                            3                             3         Blood
3                           15                            14         Blood
4                            1                             1         Blood
5                            7                             7         Blood
6                            7                             7         Blood
  visit infancy_vac biological_sex              ethnicity               race
1     3          aP         Female     Hispanic or Latino More Than One Race
2     3          wP         Female Not Hispanic or Latino              Asian
3     5          wP           Male     Hispanic or Latino More Than One Race
4     2          aP         Female     Hispanic or Latino              White
5     4          aP         Female     Hispanic or Latino More Than One Race
6     4          wP         Female Not Hispanic or Latino              Asian
  year_of_birth date_of_boost      dataset
1    1998-01-01    2016-11-07 2020_dataset
2    1989-01-01    2016-09-26 2020_dataset
3    1990-01-01    2016-10-10 2020_dataset
4    1997-01-01    2016-10-24 2020_dataset
5    1998-01-01    2016-11-07 2020_dataset
6    1989-01-01    2016-09-26 2020_dataset
```

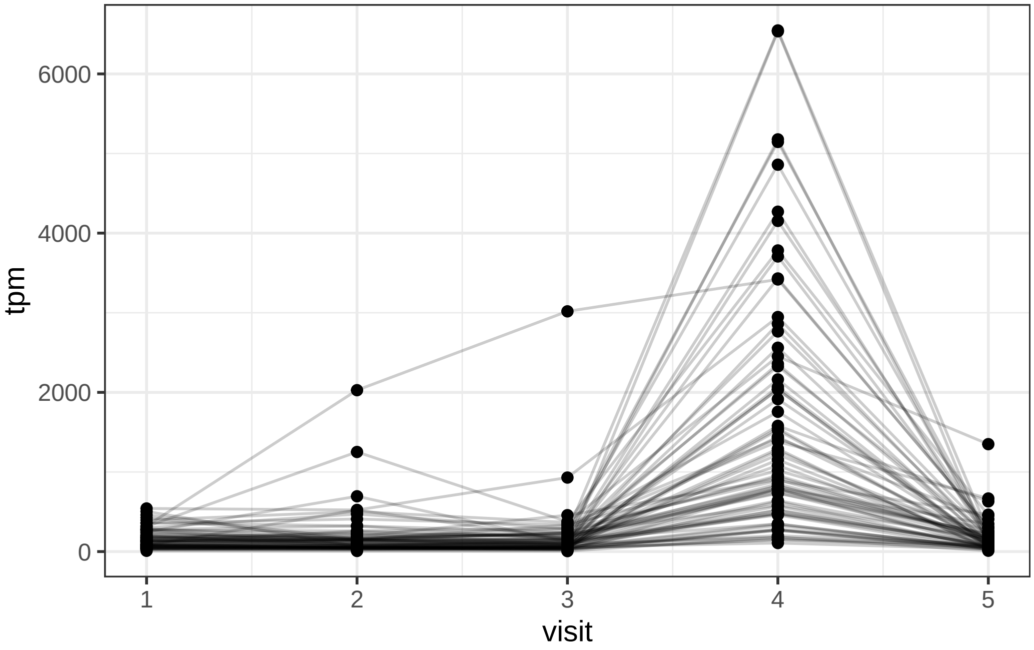
```
dim(ssrna)
```

```
[1] 360  16
```

Year of birth not age because age would be different for different visits.

Make a plot of the time course of gene expression for IGHG1 gene (i.e. a plot of visit vs. tpm).

```
ggplot(ssrna) +
  aes(visit, tpm, group=subject_id) +
  geom_point() +
  geom_line(alpha=0.2) +
  theme_bw()
```



Expression peaks at visit 4, before Ab titer peak at visit 5 or 6.