Class 19

Arturo Agramont

Pertussis Cases by Year

- **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.

We will use datapasta to copy the table from the website as a data frame.

```
cdc = data.frame(
                          Year = c(1922L,
                               1923L,1924L,1925L,1926L,1927L,1928L,
                               1929L,1930L,1931L,1932L,1933L,1934L,1935L,
                               1936L,1937L,1938L,1939L,1940L,1941L,
                               1942L,1943L,1944L,1945L,1946L,1947L,1948L,
                               1949L,1950L,1951L,1952L,1953L,1954L,
                               1955L,1956L,1957L,1958L,1959L,1960L,
                               1961L,1962L,1963L,1964L,1965L,1966L,1967L,
                               1968L,1969L,1970L,1971L,1972L,1973L,
                               1974L,1975L,1976L,1977L,1978L,1979L,1980L,
                               1981L,1982L,1983L,1984L,1985L,1986L,
                               1987L,1988L,1989L,1990L,1991L,1992L,1993L,
                               1994L,1995L,1996L,1997L,1998L,1999L,
                               2000L,2001L,2002L,2003L,2004L,2005L,
                               2006L,2007L,2008L,2009L,2010L,2011L,2012L,
                               2013L,2014L,2015L,2016L,2017L,2018L,
                               2019L,2020L,2021L),
            No..Reported.Pertussis.Cases = c(107473,
                               164191,165418,152003,202210,181411,
                               161799,197371,166914,172559,215343,179135,
                               265269,180518,147237,214652,227319,103188,
                               183866,222202,191383,191890,109873,
                               133792,109860,156517,74715,69479,120718,
```

```
                                        68687,45030,37129,60886,62786,31732,28295,
                                        32148,40005,14809,11468,17749,17135,
                                        13005,6799,7717,9718,4810,3285,4249,
                                        3036,3287,1759,2402,1738,1010,2177,2063,
                                        1623,1730,1248,1895,2463,2276,3589,
                                        4195,2823,3450,4157,4570,2719,4083,6586,
                                        4617,5137,7796,6564,7405,7298,7867,
                                        7580,9771,11647,25827,25616,15632,10454,
                                        13278,16858,27550,18719,48277,28639,
                                        32971,20762,17972,18975,15609,18617,6124,
                                        2116)
)
```

We will now make a plot of the data frame.

```
library(ggplot2)
ggplot(cdc) +
  aes(Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +
  labs(x = "Year", y = "Number of cases")
```
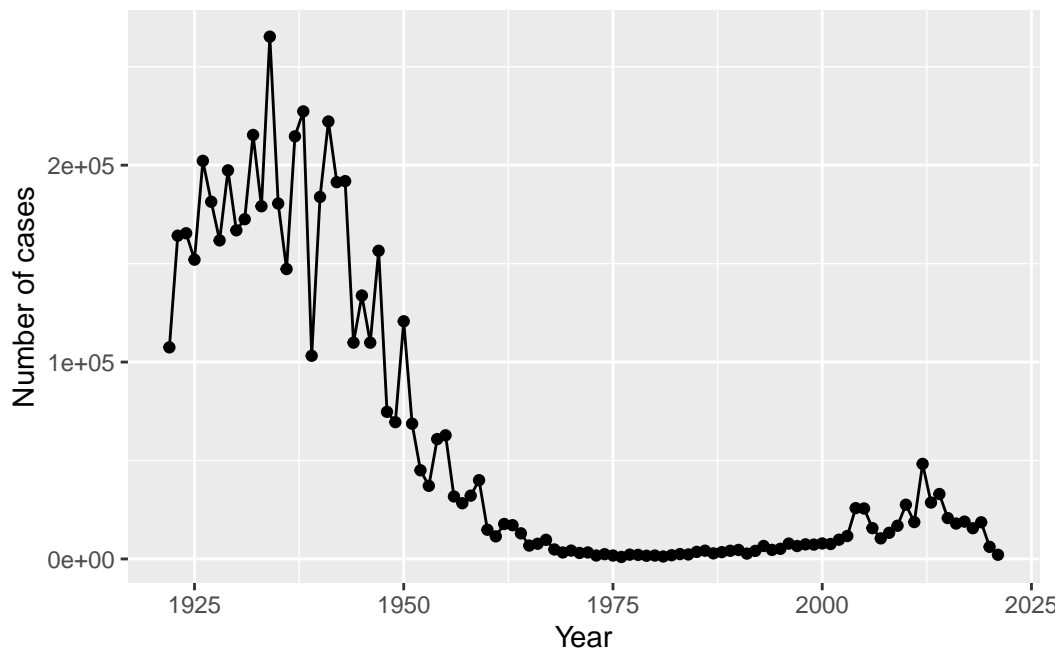
2. A tale of two vaccines (wP & aP)

- **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_point() +
  geom_line() +
  labs(x = "Year", y = "Number of cases") +
  geom_vline(xintercept = 1946,linetype = 2, color = "blue") +
  geom_vline(xintercept = 1996, linetype = 2, color = "red") +
  geom_text(aes(x=1950,y = 250000, label = "wP")) +
  geom_text(aes(x=2000,y = 250000, label = "aP"))
```
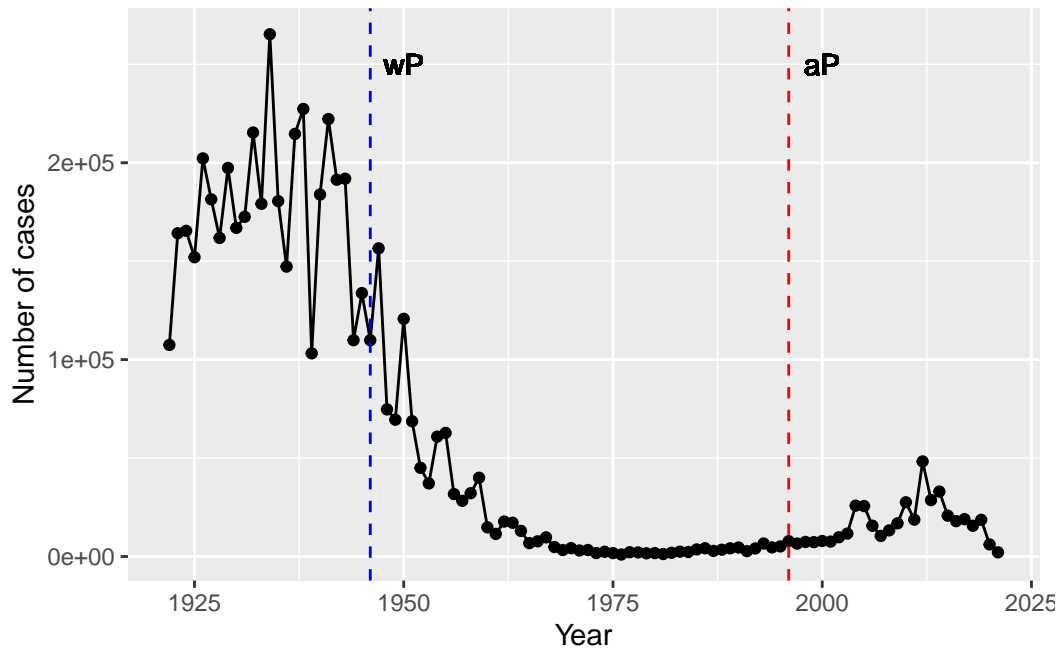

- **Q3.** Describe what happened after the introduction of the aP vaccine? Do you have a possible explanation for the observed trend?

After the introduction of the AP vaccine, the number of cases of Pertussis increased by year. A possible explanation for this is bacterial evolution, the bacteria evolving over time to become more resistant to immune system and effects of vaccines.

3. Exploring CMI-PB data

We will use jsonlite package in order to read data from the CMI-PB API database. We will then explore this data

```
library(jsonlite)
subject <- read_json("https://www.cmi-pb.org/api/subject", simplifyVector = TRUE)
```

```
head(subject, 3)
```

```
  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
  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
```

- **Q4.** How many aP and wP infancy vaccinated subjects are in the dataset?

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

```
aP wP
47 49
```

There are 47 aP vaccinated and 49 vaccinated subjects.

- **Q5.** How many Male and Female subjects/patients are in the dataset?

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

```
Female    Male
    66      30
```

There are 66 females and 30 males

- **Q6.** What is the breakdown of race and biological sex (e.g. number of Asian females, White males etc…)?

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

```
       American Indian/Alaska Native Asian Black or African American
Female                            0    18                          2
Male                              1     9                          0

       More Than One Race Native Hawaiian or Other Pacific Islander
Female                  8                                         1
Male                    2                                         1

       Unknown or Not Reported White
Female                      10    27
Male                         4    13
```

We will use lubridate to work with dates from this data

```r
library(lubridate)
```

```
Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

```r
today()
```

```
[1] "2023-06-16"
```

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

```
Time difference of 8567 days
```

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

```
[1] 23.45517
```

- **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?

```r
subject$age = today() - ymd(subject$year_of_birth)
```

```r
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
```

```r
ap <- subject %>% filter(infancy_vac == "aP")

round( summary( time_length( ap$age, "years" ) ) )
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     23      25      26      26      26      27
```

```r
wp <- subject %>% filter(infancy_vac == "wP")
round( summary( time_length( wp$age, "years" ) ) )
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     28      32      35      37      40      55
```

```r
t.test(ap$age, wp$age)
```

```
    Welch Two Sample t-test

data:  ap$age and wp$age
t = -12.092 days, df = 51.082, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4618.534 days -3303.337 days
sample estimates:
Time differences in days
mean of x mean of y
 9413.574 13374.510
```

These results are significantly different.

- **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?

```
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`.
```
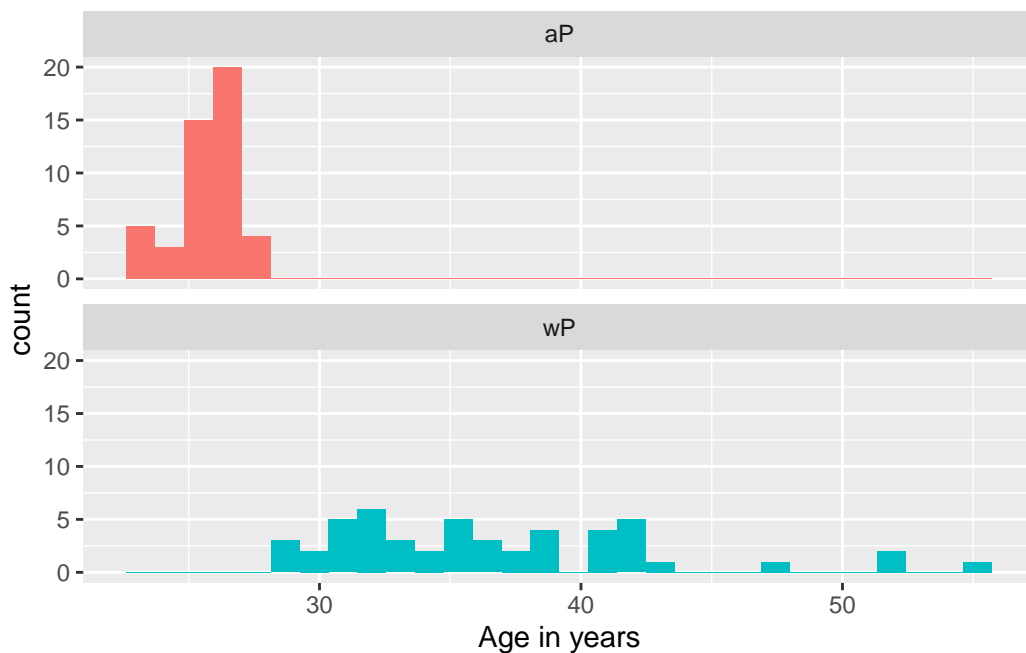
The graph shows clear differences between the 2 groups, therefore they are most likely significantly different.

```r
specimen <- read_json("https://www.cmi-pb.org/api/specimen", simplifyVector = TRUE)
titer <- read_json("https://www.cmi-pb.org/api/ab_titer", simplifyVector = TRUE)
```

Now we will join specimen and subject together

```r
meta <- inner_join(specimen, subject)
```

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

```r
dim(meta)
```

```
[1] 729  14
```

```r
head(meta)
```

```
  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 infancy_vac biological_sex
1                             0         Blood     1          wP         Female
2                           736         Blood    10          wP         Female
3                             1         Blood     2          wP         Female
4                             3         Blood     3          wP         Female
5                             7         Blood     4          wP         Female
6                            14         Blood     5          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
1 13680 days
2 13680 days
3 13680 days
4 13680 days
5 13680 days
6 13680 days
```

- **Q11.** 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.

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

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

```
dim(abdata)
```

```
[1] 32675    21
```

- **Q12.** How many specimens (i.e. entries in `abdata`) do we have for each `isotype`?

9

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

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

- **Q13.** What do you notice about the number of visit 8 specimens compared to other visits?

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

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

The number of visit 8 specimens are much lower compared to the other visits

4. Examine IgG1 Ab titer levels

We will use filter to exclude visit 8 entries

```r
ig1 <- abdata %>% filter(isotype == "IgG1", visit!=8)
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
        age
1 13680 days
2 13680 days
3 13680 days
4 13680 days
5 13680 days
6 13680 days
```
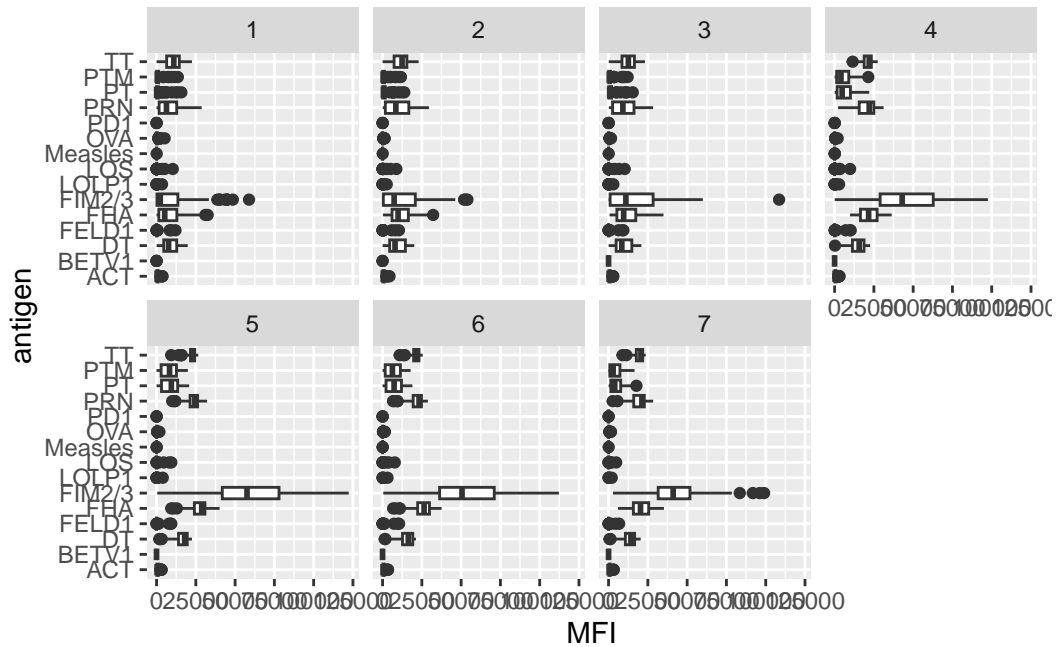
- **Q14.** Complete the following code to make a summary boxplot of Ab titer levels (MFI) for all antigens:

We are going to make a v=boxplot of the Ab titer levels

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

- **Q15.** What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

FIM2/3 show a large difference in levels of IgG1 antibody titers.

```r
ggplot(ig1) +
  aes(MFI, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  theme_bw()
```
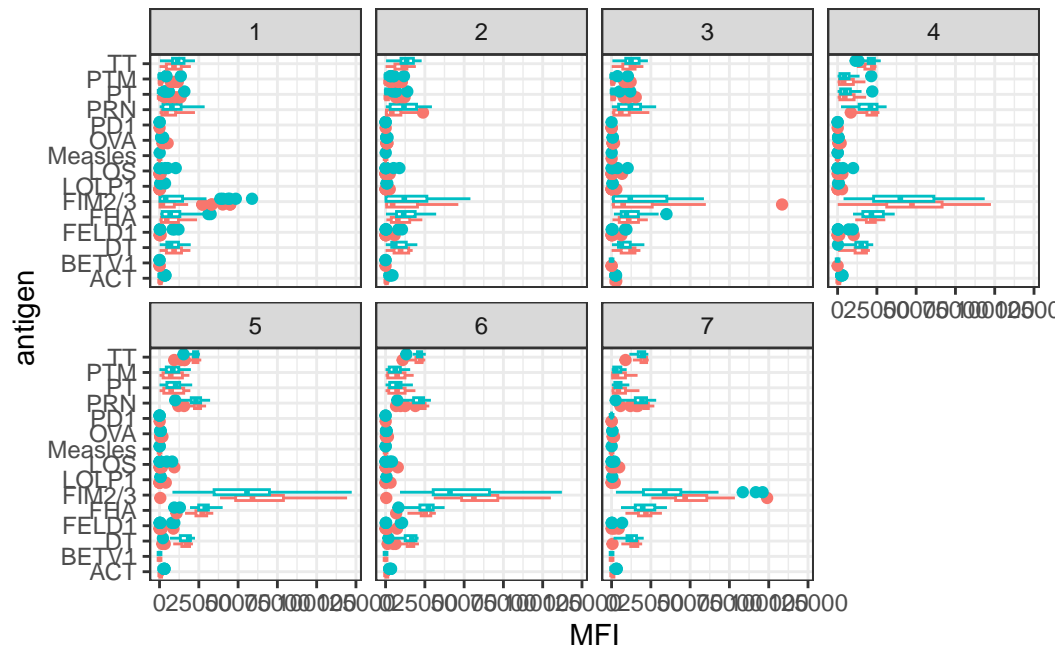
```
ggplot(ig1) +
  aes(MFI, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```
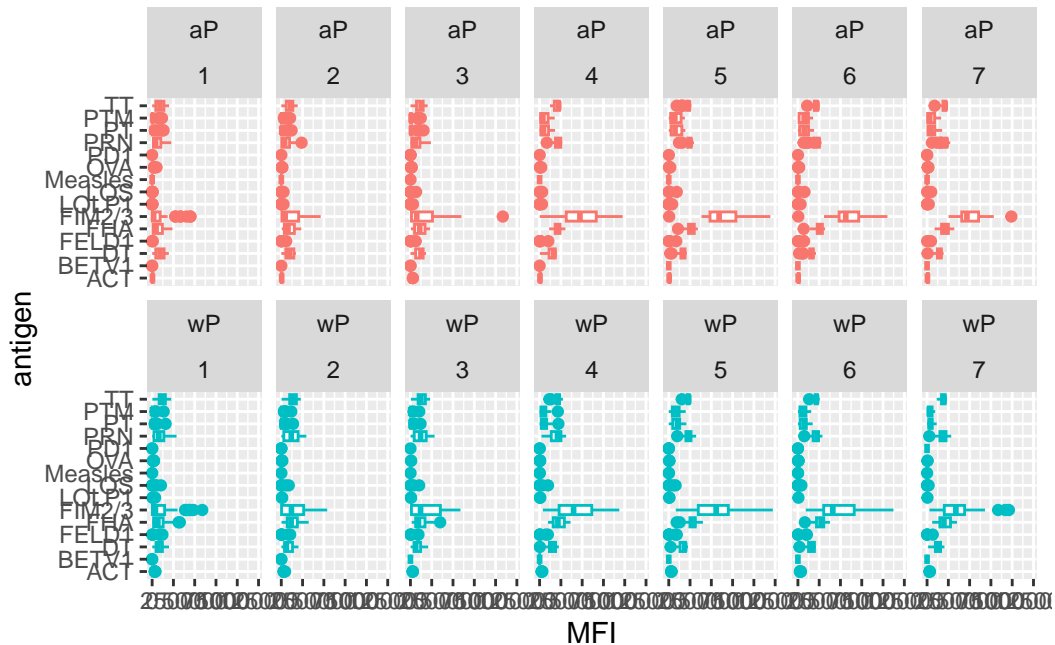
- **Q16.** Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen (**"Measles"**, that is not in our vaccines) and a clear antigen of interest (**"FIM2/3"**, extra-cellular fimbriae proteins from *B. pertussis* that participate in substrate attachment).

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

MFI
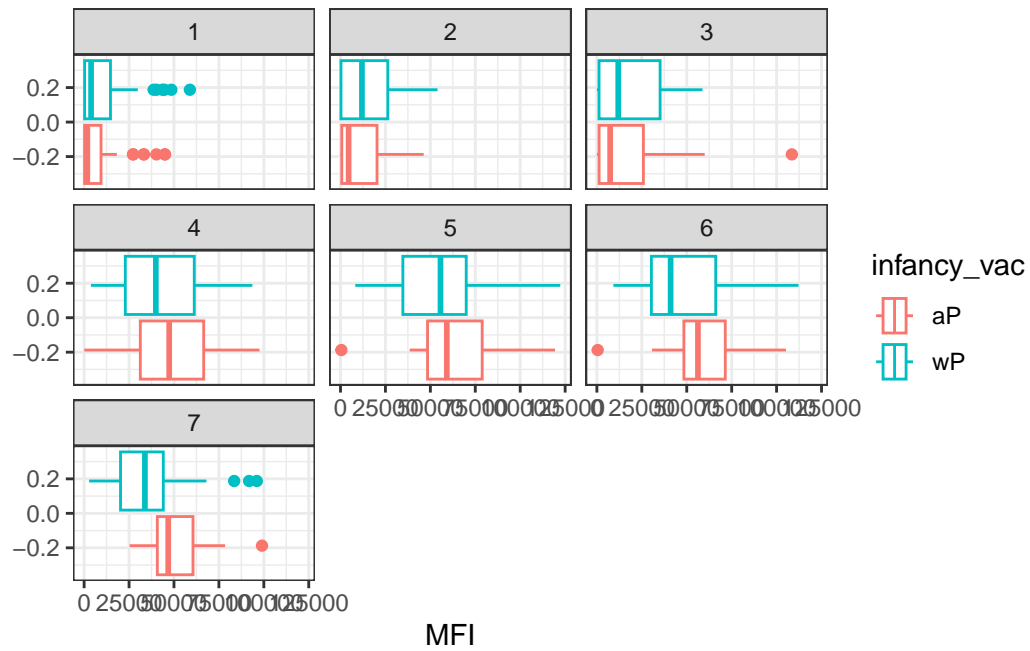
```r
filter(ig1, antigen=="FHA") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = T) +
  facet_wrap(vars(visit)) +
  theme_bw()
```
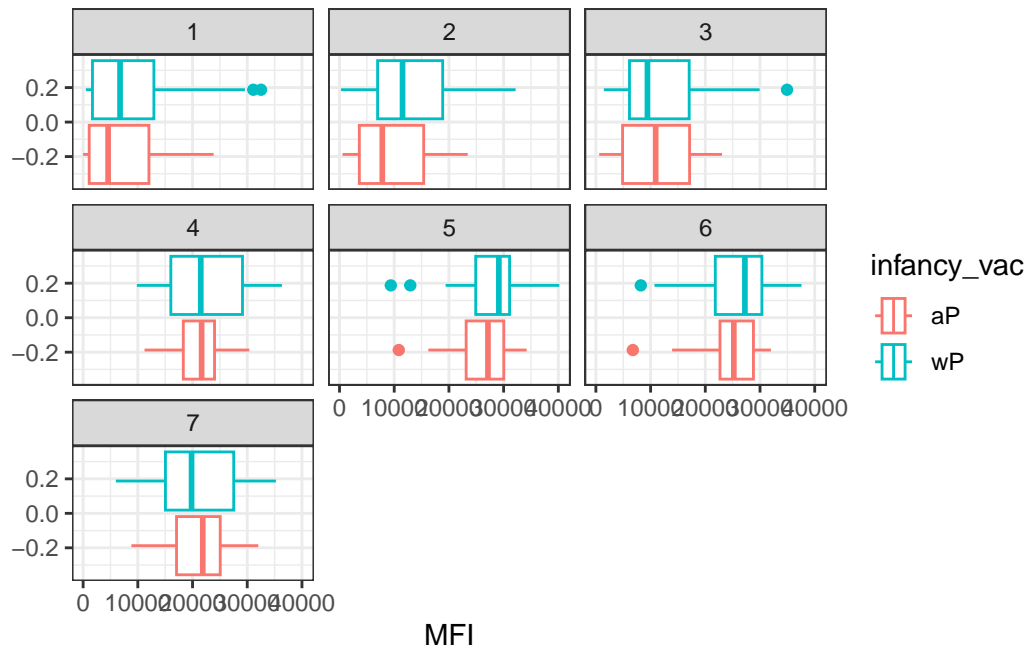
- **Q17.** What do you notice about these two antigens time courses and the FIM2/3 data in particular?

Peak at 5, with a decline. FHA, and FIM2/3 both have an increase over time.

- **Q18.** Do you see any clear difference in aP vs. wP responses?

wP increases more at first but then aP has a greater increase afterward.

5. Obtaining CMI-PB RNASeq data

We are going to read RNA-seq data

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

rna <- read_json(url, simplifyVector = TRUE)
```
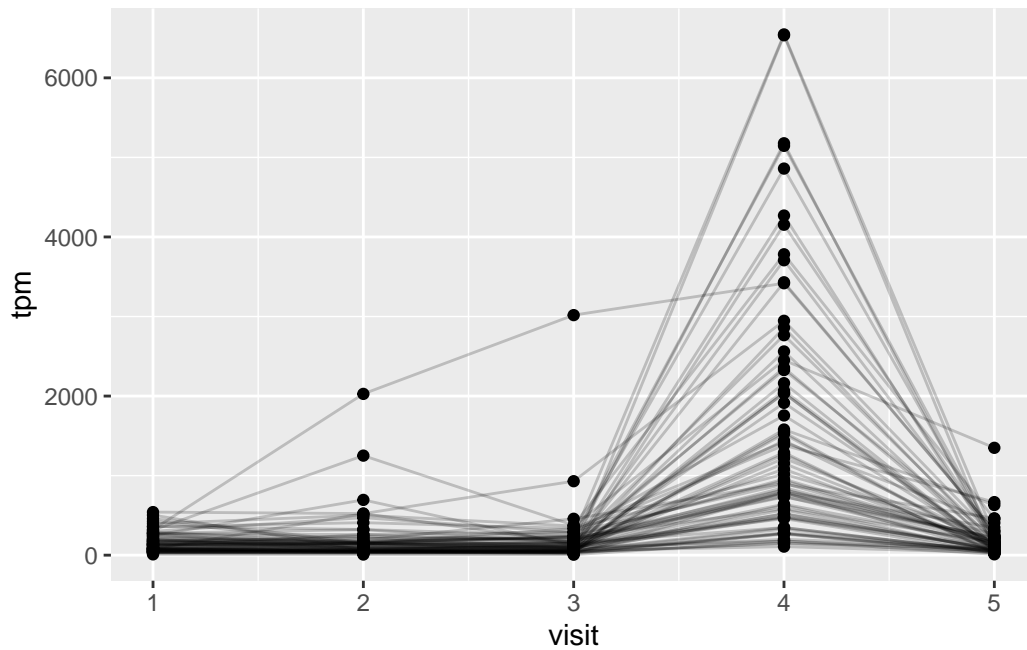
We will join RNA data with meta data

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

Joining with `by = join_by(specimen_id)`

- **Q19.** 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)
```


- **Q20.**: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

Visit 4 had the highest amount of expression. With a sharp decrease afterward.

- **Q21.** Does this pattern in time match the trend of antibody titer data? If not, why not?

The titer data demonstrated that the highest expression occurred during visit 5, but in this instance visit 4 had the highest The cells expressed the gene earlier on than when the antibodies were at their greatest.