Class 17:

Katherine Lim (A15900881)

1. Investigating pertussis cases by year

Pertussis or whooping cough is a highly contagious lung infection caused by the bacteria *B. pertussis*.

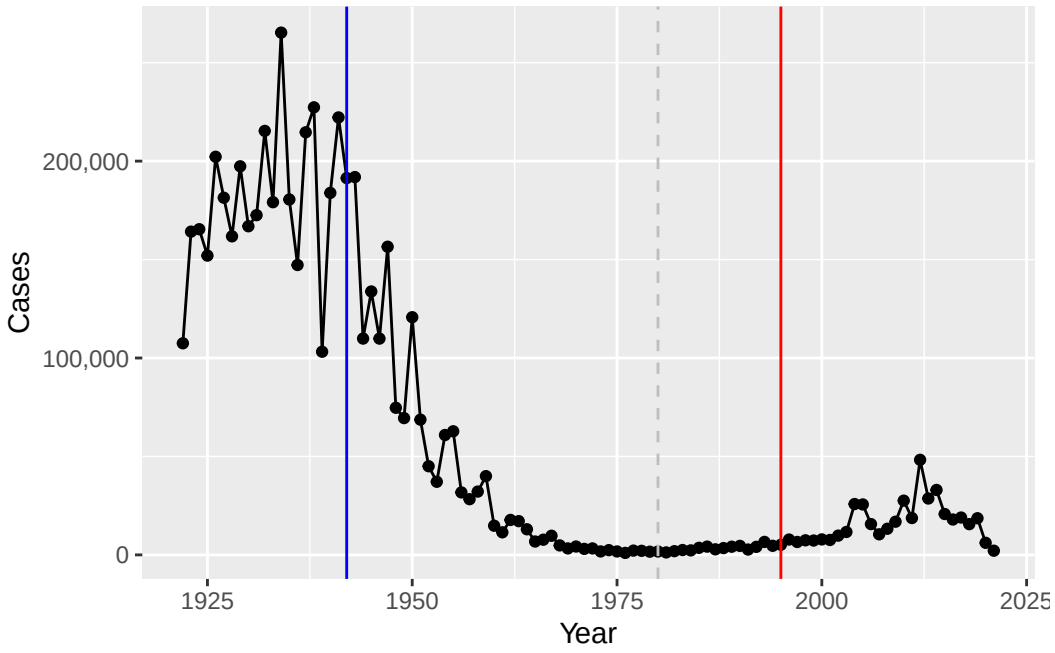
The CDC tracks reported cases in the US since the 1920s.

> 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.

```
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),
```

```
        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)
       )
```

```
library(ggplot2)

ggplot(cdc) +
  aes(Year, Cases) +
  geom_point() +
  geom_line() +
  labs(x = "Year", Y = "Cases") +
  geom_vline(xintercept = 1942, color = "blue") +
  geom_vline(xintercept = 1980, color = "grey", linetype = 2) +
  geom_vline(xintercept = 1995, color = "red") +
  scale_y_continuous(labels = scales::comma)
```

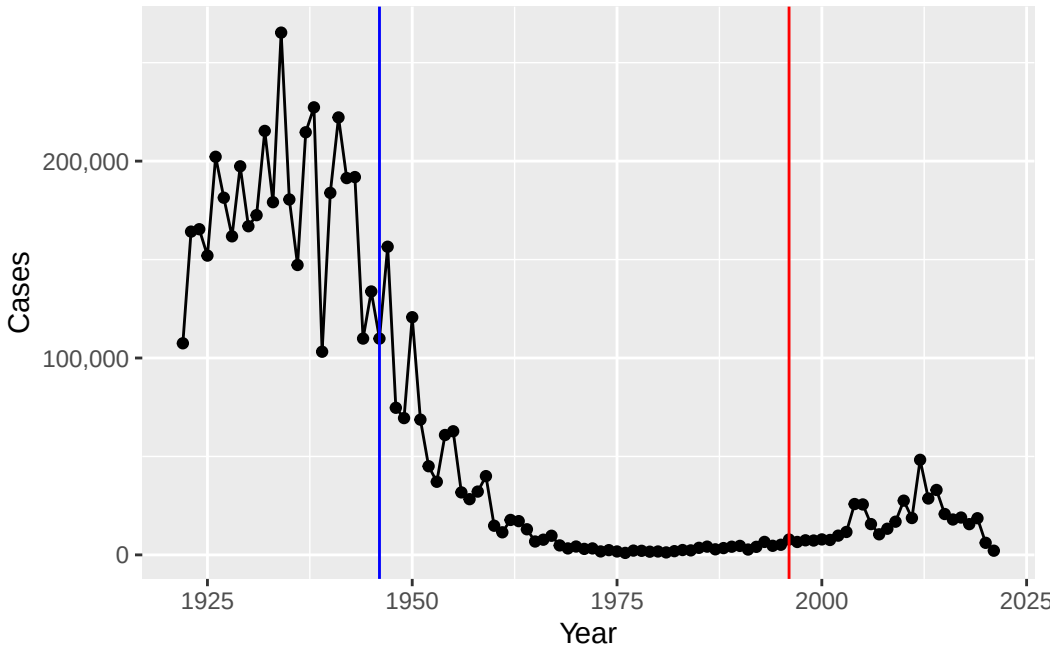


The first big "whole-cell" pertussis vaccine program started in 1942.

2. A tale of two vaccines

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.

```
library(ggplot2)

ggplot(cdc) +
  aes(Year, Cases) +
  geom_point() +
  geom_line() +
  labs(x = "Year", Y = "Cases") +
  geom_vline(xintercept = 1946, color = "blue") +
  geom_vline(xintercept = 1996, color = "red") +
  scale_y_continuous(labels = scales::comma)
```



Something big is happening with pertussis and big outbreaks are once again a major public health concern.

One of the main hypotheses for the increasing case numbers is waning vaccine efficiency with the newer aP vaccine.

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

There is a slight increase in cases following the switch to the aP vaccine.

3. Exploring CMI-PB data

Enter the CMI-PB project, which is studying this problem on a large scale. Let's see what data they have.

Their data is available in JSON format ("key:value" pair style). We will use the "jsonlight" package to read their 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 may aP and wP infancy vaccinated subjects are in the dataset?

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

```
aP wP
47 49
```

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

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

```
Female   Male
    66     30
```

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                                     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
```

Q7. 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
```

```
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
```

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

# (i) the average age of wP individuals
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
```

```
# (ii) the average age of aP individuals
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
```

```
# (iii) are they significantly different
t_test <- t.test(wp$age, ap$age)
print(t_test)
```

```
    Welch Two Sample t-test

data:  wp$age and ap$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:
 3303.337 days 4618.534 days
sample estimates:
Time differences in days
mean of x mean of y
13364.510  9403.574
```

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)
```

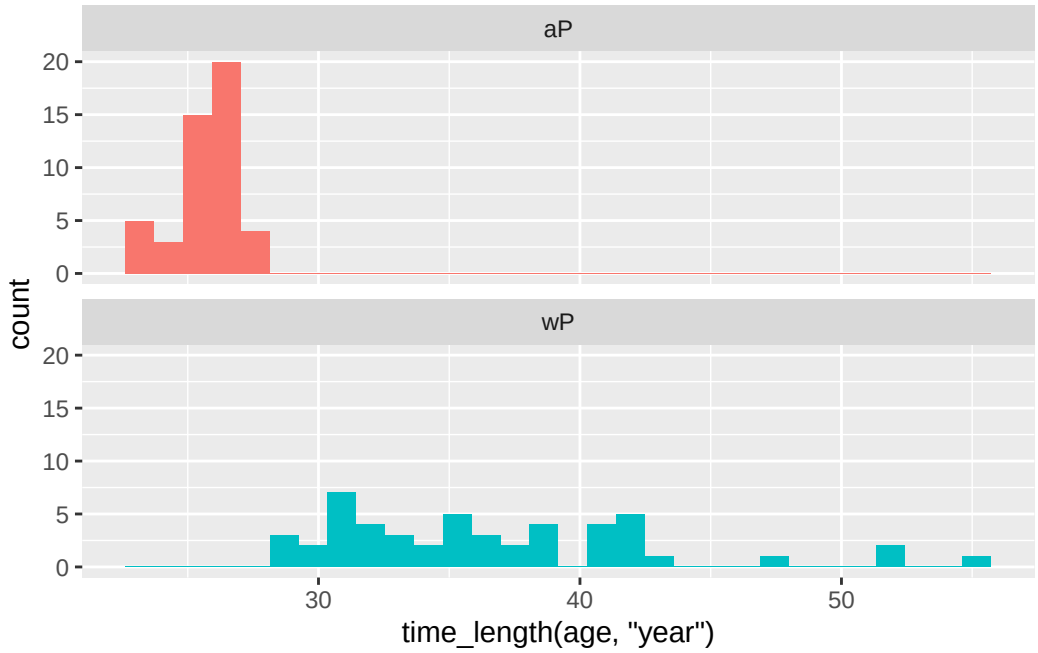
```
[1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481
```

Q9. With the help of a faceted boxplot (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)
```

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



Yes.

Now let's read some more database tables from CMI-PB.

```
specimen <- read_json("http://cmi-pb.org/api/specimen", 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        age
```

```
1    1986-01-01    2016-09-12 2020_dataset 13670 days
2    1968-01-01    2019-01-28 2020_dataset 20245 days
3    1983-01-01    2016-10-10 2020_dataset 14766 days
4    1988-01-01    2016-08-29 2020_dataset 12940 days
5    1991-01-01    2016-08-29 2020_dataset 11844 days
6    1988-01-01    2016-10-10 2020_dataset 12940 days
```

I want to join" the `subject` and `specimen` tables together. We can use the `dplyr` package for this.

> 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.

```
library(dplyr)

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 13670 days           1
2    1986-01-01    2016-09-12 2020_dataset 13670 days           2
3    1986-01-01    2016-09-12 2020_dataset 13670 days           3
4    1986-01-01    2016-09-12 2020_dataset 13670 days           4
5    1986-01-01    2016-09-12 2020_dataset 13670 days           5
6    1986-01-01    2016-09-12 2020_dataset 13670 days           6
  actual_day_relative_to_boost planned_day_relative_to_boost specimen_type
1                           -3                             0         Blood
2                          736                           736         Blood
3                            1                             1         Blood
```

```
4                            3                              3         Blood
5                            7                              7         Blood
6                           11                             14         Blood
  visit
1     1
2    10
3     2
4     3
5     4
6     5
```

Q10. Now using the same procedure join ab with meta data so we can further analyze this data in terms of time of visit aP/wP, male/female etc.

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

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

```
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        age specimen_id
1    1986-01-01    2016-09-12 2020_dataset 13670 days           1
2    1986-01-01    2016-09-12 2020_dataset 13670 days           1
3    1986-01-01    2016-09-12 2020_dataset 13670 days           1
4    1986-01-01    2016-09-12 2020_dataset 13670 days           1
5    1986-01-01    2016-09-12 2020_dataset 13670 days           1
6    1986-01-01    2016-09-12 2020_dataset 13670 days           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
```

Q11. How many specimens (i.e. entries in abdata) do we have for each isotype?

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

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

Q12. What do you notice about the number of visit 8 specimens compared to other visits?

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

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

There are way less visit 8 specimens because the project is still ongoing and we do not have all the data for those individuals yet.

4. Examine IgG1 Ab titer levels

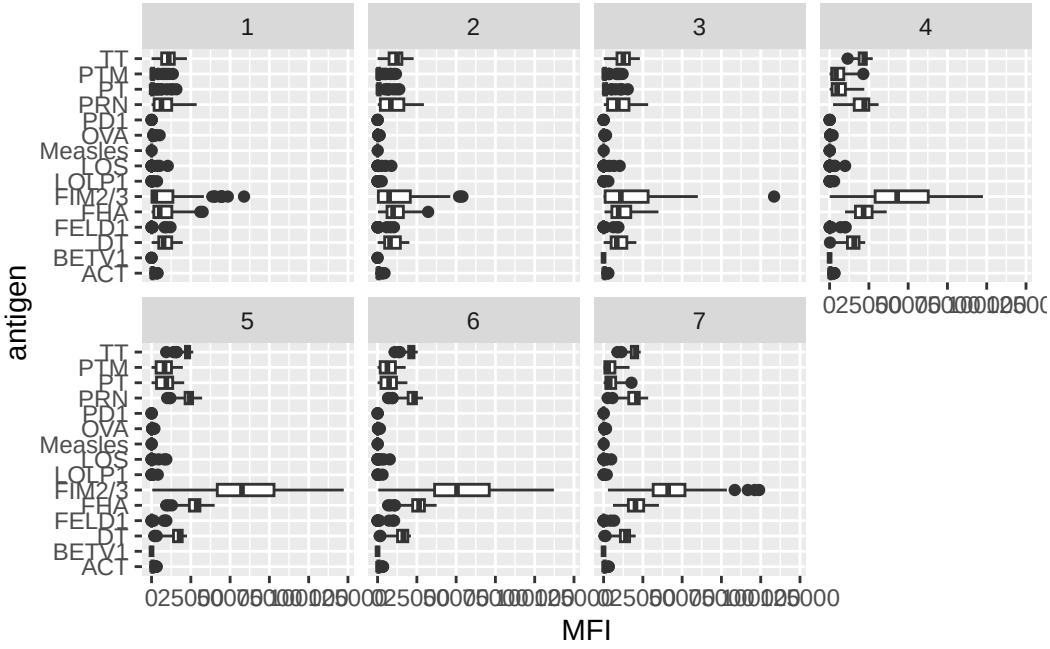
We will use the `filter()` function from dplyr to focus on just IgG1 isotype and visits 1 to 7.

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

```
  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 13670 days           1
2    1986-01-01    2016-09-12 2020_dataset 13670 days           1
3    1986-01-01    2016-09-12 2020_dataset 13670 days           1
4    1986-01-01    2016-09-12 2020_dataset 13670 days           1
5    1986-01-01    2016-09-12 2020_dataset 13670 days           1
6    1986-01-01    2016-09-12 2020_dataset 13670 days           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    IgG1                TRUE     ACT 274.355068      0.6928058 IU/ML
2     1    IgG1                TRUE     LOS  10.974026      2.1645083 IU/ML
3     1    IgG1                TRUE   FELD1   1.448796      0.8080941 IU/ML
4     1    IgG1                TRUE   BETV1   0.100000      1.0000000 IU/ML
5     1    IgG1                TRUE   LOLP1   0.100000      1.0000000 IU/ML
6     1    IgG1                TRUE Measles  36.277417      1.6638332 IU/ML
  lower_limit_of_detection
1                 3.848750
2                 4.357917
3                 2.699944
4                 1.734784
5                 2.550606
6                 4.438966
```

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

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

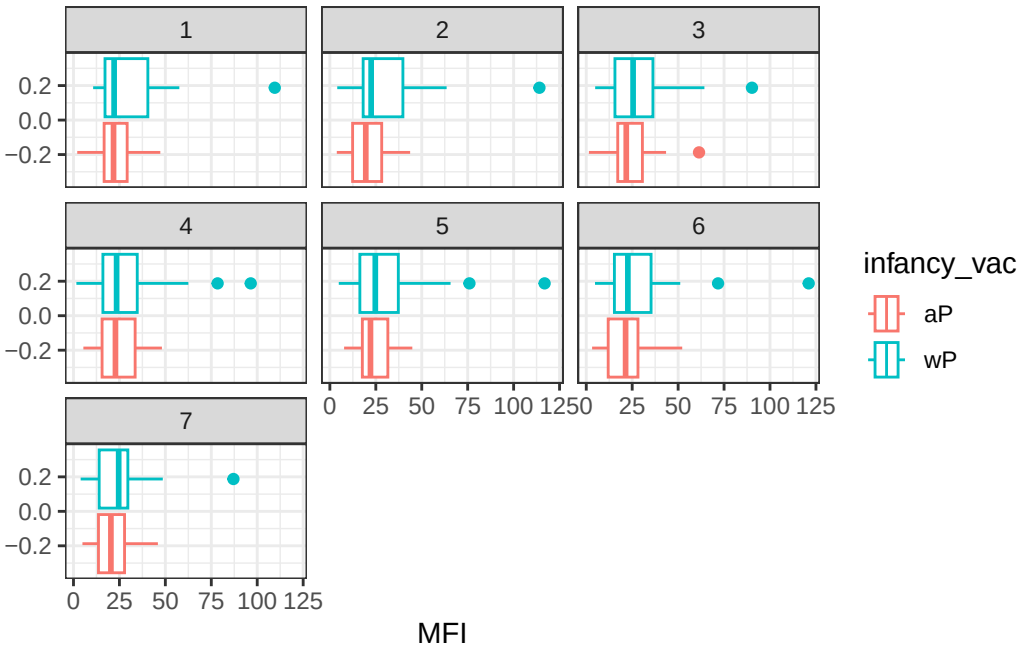


Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

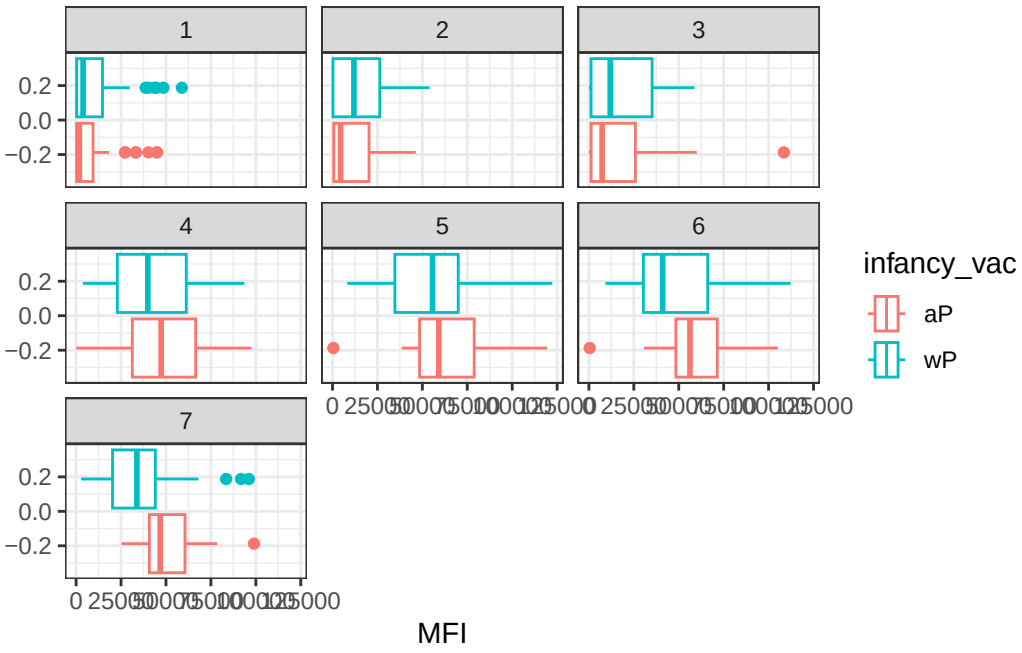
The FIM 2/3 (fimbrial protein), PT (pertussis toxin), FHA (filamentous hemagglitinin).

Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each.

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




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



Q16. What do you notice about these two antigens time course and the FIM2/3 data in particular?

FIM2/3 levels rise over time and exceed those of Measles. They also appear to peak at visit 5 and then decline.

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

wP and aP responses follow a similar trend.

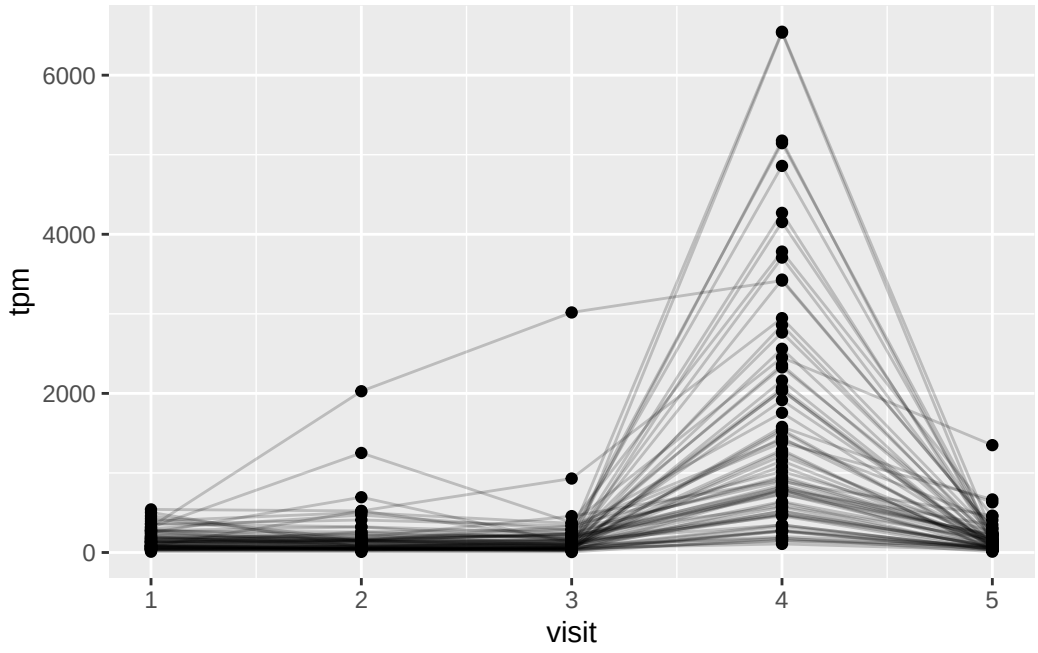
5. Obtaining CMI-PB RNASeq data

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

rna <- read_json(url, simplifyVector = TRUE)

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

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

Q18. 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)
```



Q19. What do you notice about the expression of this gene?

The expression of this gene is at its maximum at visit 4.