Class17: Investigating Pertussis

Investigating pertussis cases by year

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

The CDC tracks reported cases in the U.S. since the 1920s

> Q1. Read the CDC data into a dataframe using datapasta

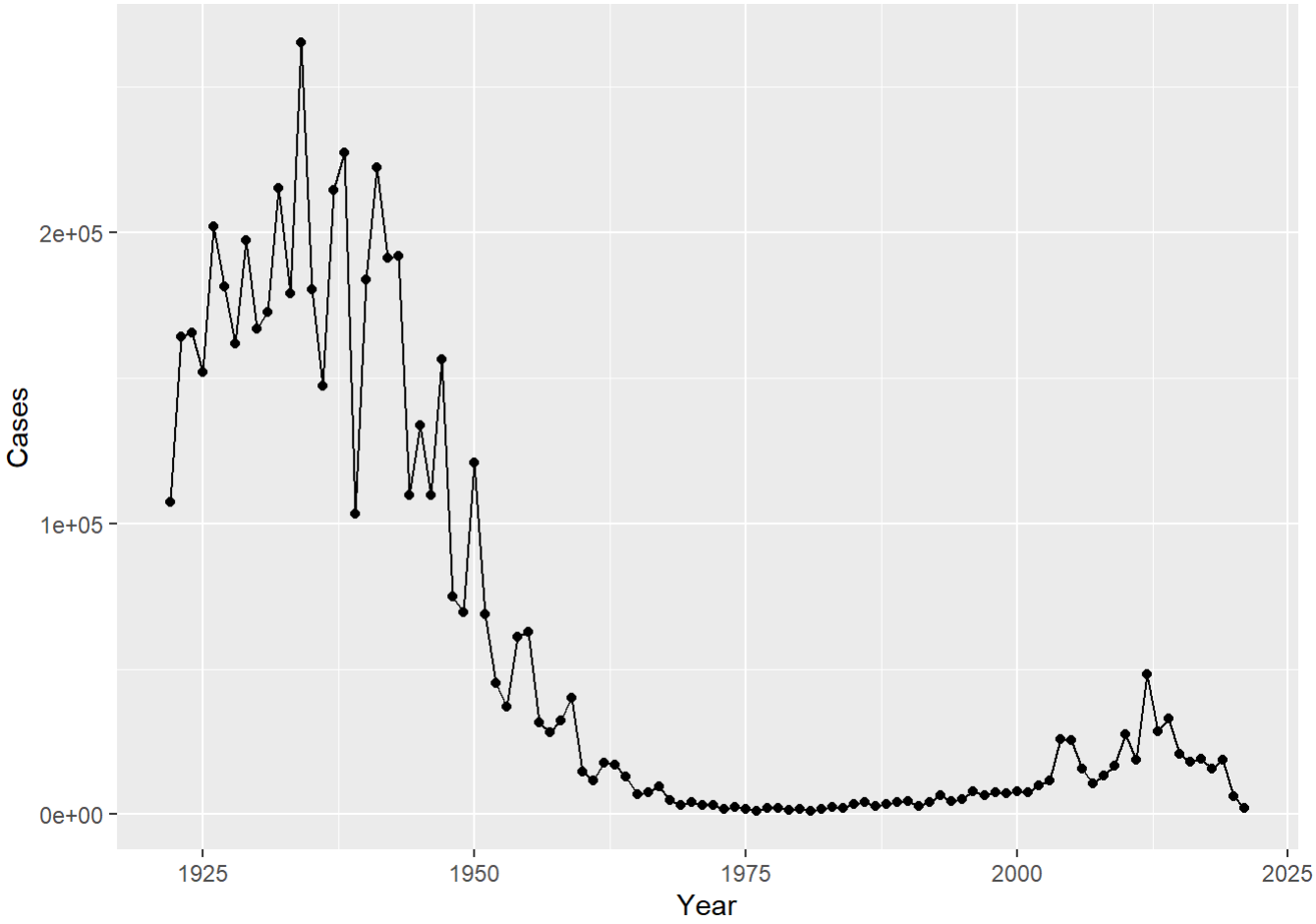
```
#step 1 install and call datapasta
library(datapasta)
#Use addin menu and select paste as dataframe (datapasta)
cdc <- data.frame(
                                 Year = c(1922L,1923L,1924L,1925L,
                                          1926L,1927L,1928L,1929L,1930L,1931L,
                                          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)
       )
```

A tale of two vaccine (wP & aP)

Plotting our cdc data

```
library(ggplot2)

pplot <- ggplot(cdc,
        aes(Year, Cases))+
  geom_line() +
  geom_point()
pplot
```

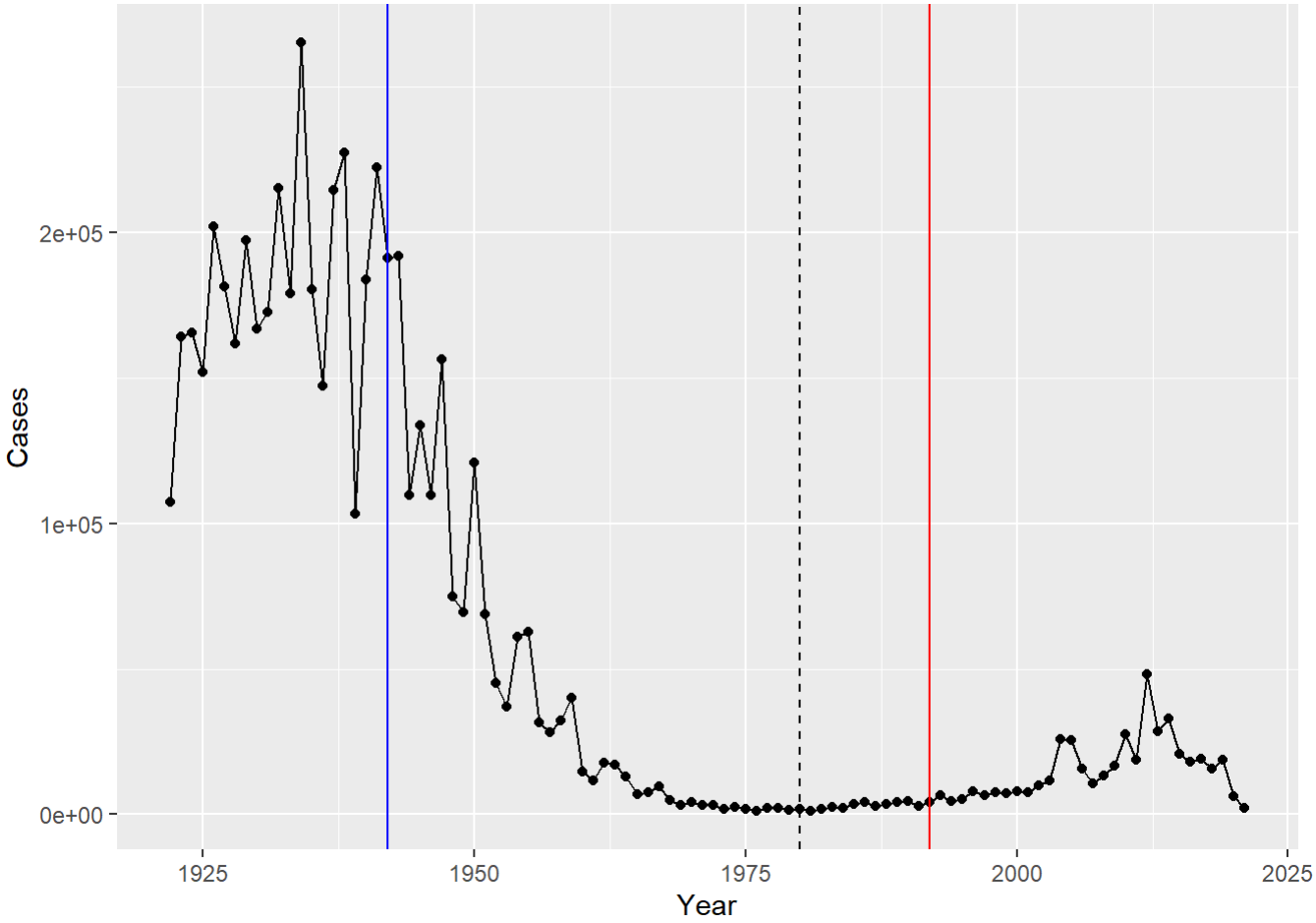


> Q2. Add a vertical line for when the vaccines were introduced

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

The vaccine doubt movement occured around the 1980s

```
pplot +
  geom_vline(xintercept = 1942, color = 'blue')+
  geom_vline(xintercept = 1980, color = 'black', linetype = 2) +
  geom_vline(xintercept = 1992, color = 'red')
```



> Q3. Describe what happened after the introduction of the aP vaccine

After it was approved in 1992 the number of cases remained relatively stable until cases began to rise around the early 2000s. A possible explanation could be the evolution of resistance to the vaccine by the bacteria. Another explanation could be that the acellular vaccine gave less immunity compared to the whole cell vaccine.

Exporing the CMI-PB data

Something bi is happening with pertussis cases and big outbreaks are once again a major public health concern! BUGGER

One of the main hypothesis for the increasing case numbers is waning vaccine efficacy with the newer aP vaccine.

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

Their data is available in JSON format ("key:value" pair style). We will use the 'jsonlite' package to read their data

```
library(jsonlite)

subject <- read_json("https://www.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
```

> Q4 How many aP and wP infancy vaccinated subjects are in the dataset?

There are 47 aP and 49 wP vaccinated

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

```
aP wP
47 49
```

> Q5 How many male and female subjects/patients are in the dataset?

There are 30 males and 66 females in the dataset

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

```
Female   Male
    66     30
```

> Q6. What is the breakdown of race and biological sex

```
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
```

Side-Note: Working with dates

```
library(lubridate)
```

```
Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

```
today()
```

```
[1] "2023-06-09"
```

```
today() - ymd("2001-06-01")
```

```
Time difference of 8043 days
```

```
time_length(today() - ymd("2001-06-01"), 'years')
```

```
[1] 22.02053
```

> Q7 Determine the average age of wP individuals, aP individuals, and if they are significantly different

The average age of wP individuals is 37 years old while the average age for aP individuals is 26 years old. The average age is statistically different according to an unpaired t.test.

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

```
#find aP age
#filter for aP subjects
ap <- subject %>% filter(infancy_vac == 'aP')
#convert days to years
round( summary (time_length(ap$age, 'years')))
```

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

```
#mean age of ap = 26 years old
```

```
#find wP age

#filter for wP subjects
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 
```

```
#mean age of wp = 37 years old

t.test(wp$age,ap$age, paired = FALSE)
```

```
    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 
13367.510  9406.574 
```

> Q8 Determine age 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)
```

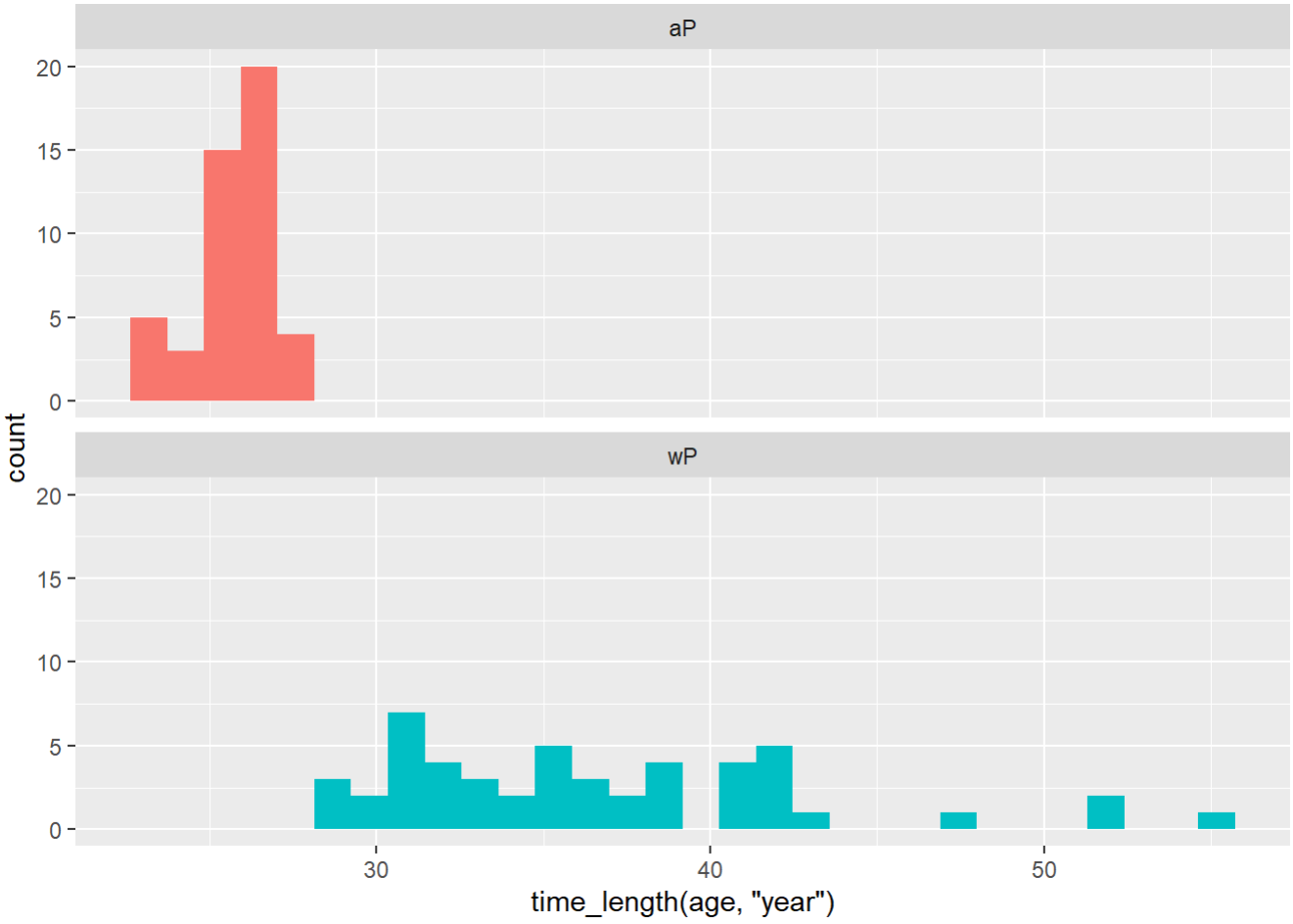
```
[1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481
```

> Q9 use a faceted boxplot to answer if you think these two groups are significantly different

I do think these two groups are significantly different, as there is very little overlap seen in the plot in addition to the t-tests I ran earlier.

```
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`.
```



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

```
#broken link from CMI-PB
#titer <- read_json('http://cmi-pb.org/api/plasma_ab_titer', simplifyVector = TRUE)
```

Joining multiple tables

Inner vs outer join - inner only keeps the data that is present in both datasets - full will keep all data, storing any missing things as NA

I want to 'join' (aka "merge"/link) the `subject` and `speciment` tables together. I will use the **dplyr** package for this.

> Q9. Complete the code to join the specimen and subject tables

```
library(dplyr)

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

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

```
dim(meta)
```

```
[1] 729  14
```

```
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 13673 days           1
2    1986-01-01    2016-09-12 2020_dataset 13673 days           2
3    1986-01-01    2016-09-12 2020_dataset 13673 days           3
4    1986-01-01    2016-09-12 2020_dataset 13673 days           4
5    1986-01-01    2016-09-12 2020_dataset 13673 days           5
6    1986-01-01    2016-09-12 2020_dataset 13673 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
```

```
ncol(specimen)
```

```
[1] 6
```

```
ncol(subject)
```

```
[1] 9
```

> Q10. Now join meta with titer data

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

```
dim(titer)
```

```
[1] 32675     8
```

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

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

```
dim(abdata)
```

```
[1] 32675    21
```

> Q11. How many specimens 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

There are very little visit 8 specimens since the project is still ongoing

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

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

Examine IgG1 Ab titer levels

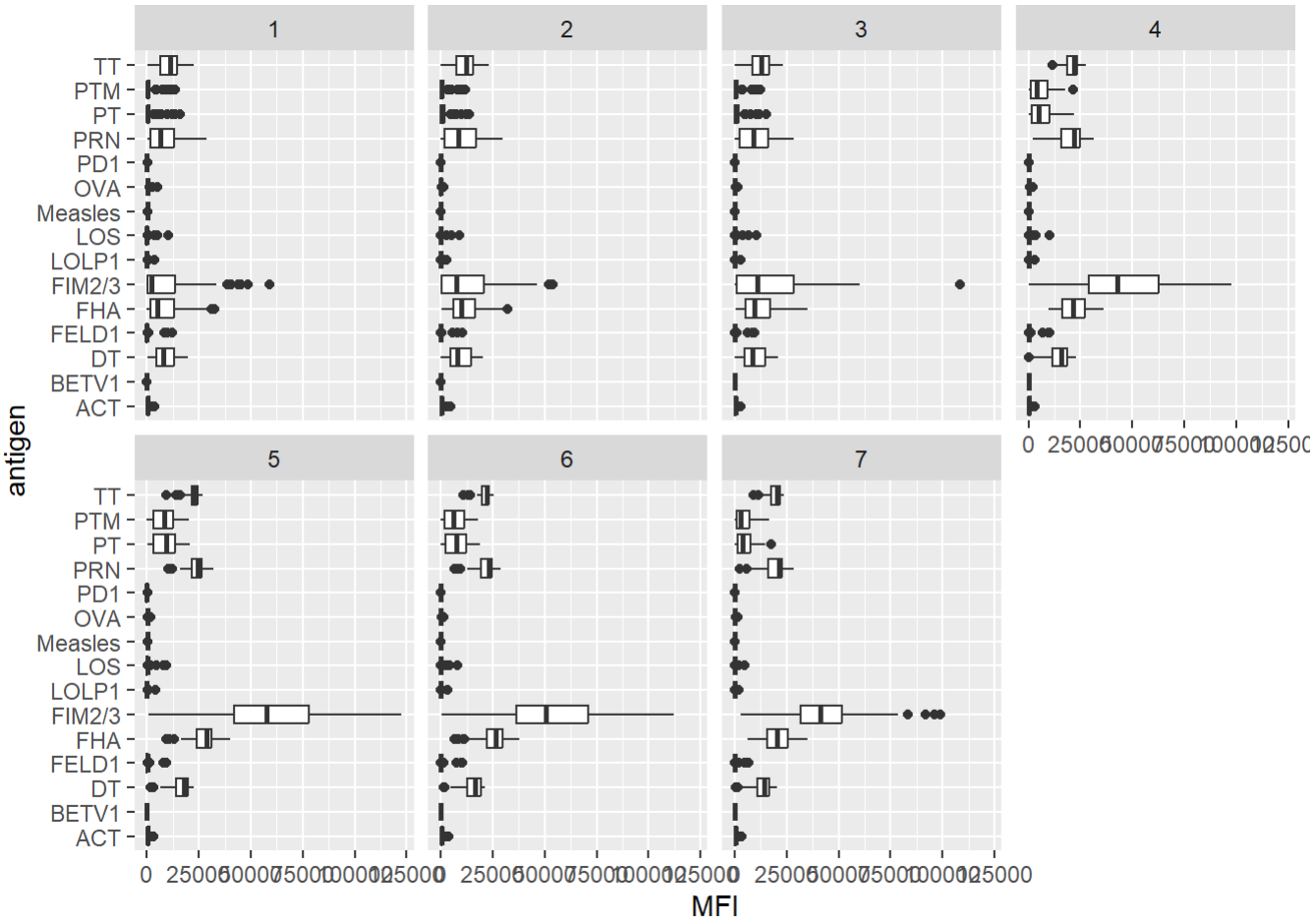
```
#filter out for IgG1 data using the filter(0 function from dplyr)

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 infancy_vac biological_sex
1 IU/ML                 3.848750          1          wP         Female
2 IU/ML                 4.357917          1          wP         Female
3 IU/ML                 2.699944          1          wP         Female
4 IU/ML                 1.734784          1          wP         Female
5 IU/ML                 2.550606          1          wP         Female
6 IU/ML                 4.438966          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 actual_day_relative_to_boost planned_day_relative_to_boost
1 13673 days                          -3                             0
2 13673 days                          -3                             0
3 13673 days                          -3                             0
4 13673 days                          -3                             0
5 13673 days                          -3                             0
6 13673 days                          -3                             0
  specimen_type visit
1         Blood     1
2         Blood     1
3         Blood     1
4         Blood     1
5         Blood     1
6         Blood     1
```

> Q13 Make a summary boxplot of Ab titer levels for all antigens

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



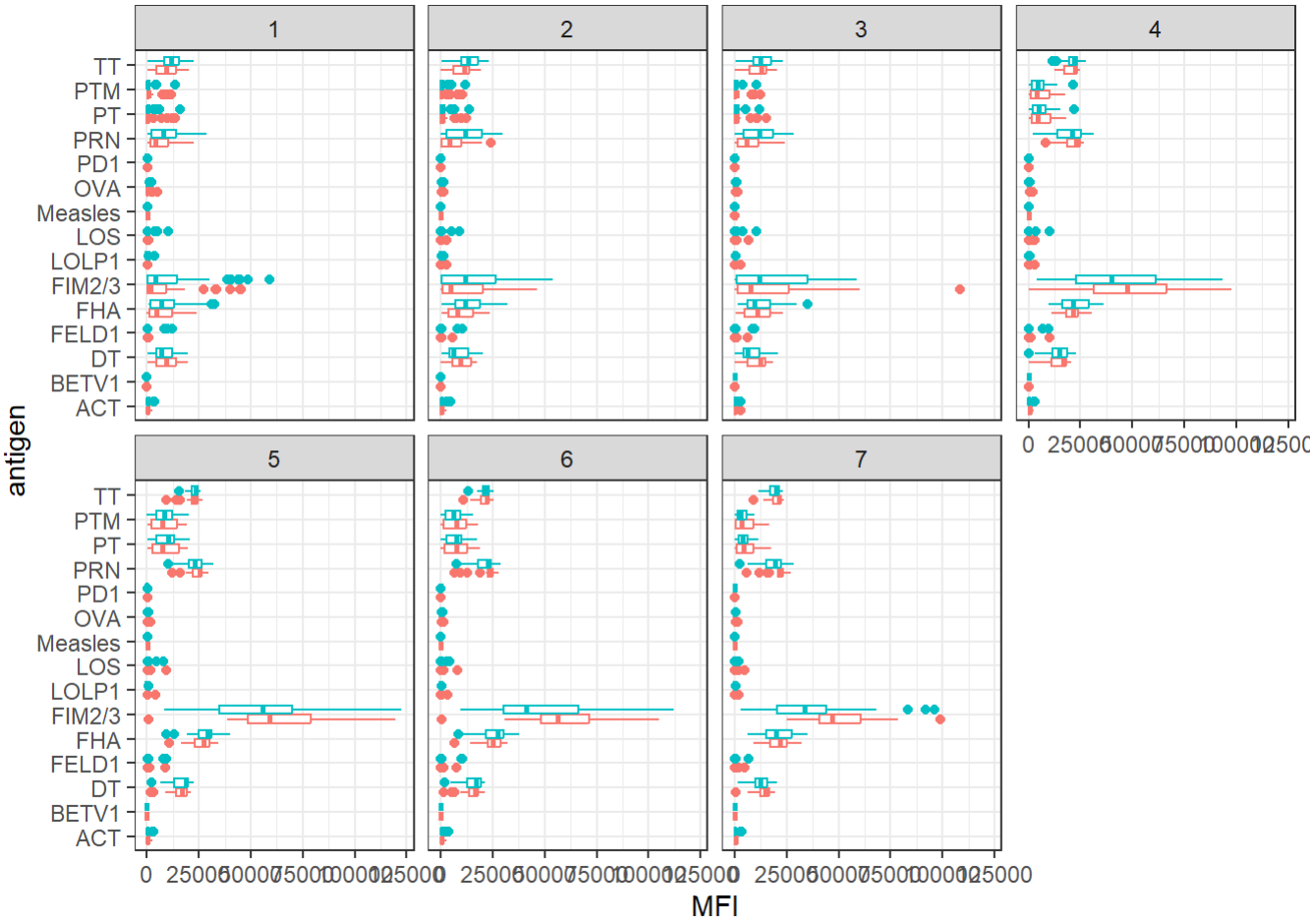
> Q14 What antigens show differences in the level of IgG1 antibody titers recognizing them. Why these and not others

FIMM23 are related to bacteria pilus and cell adhesion which is important

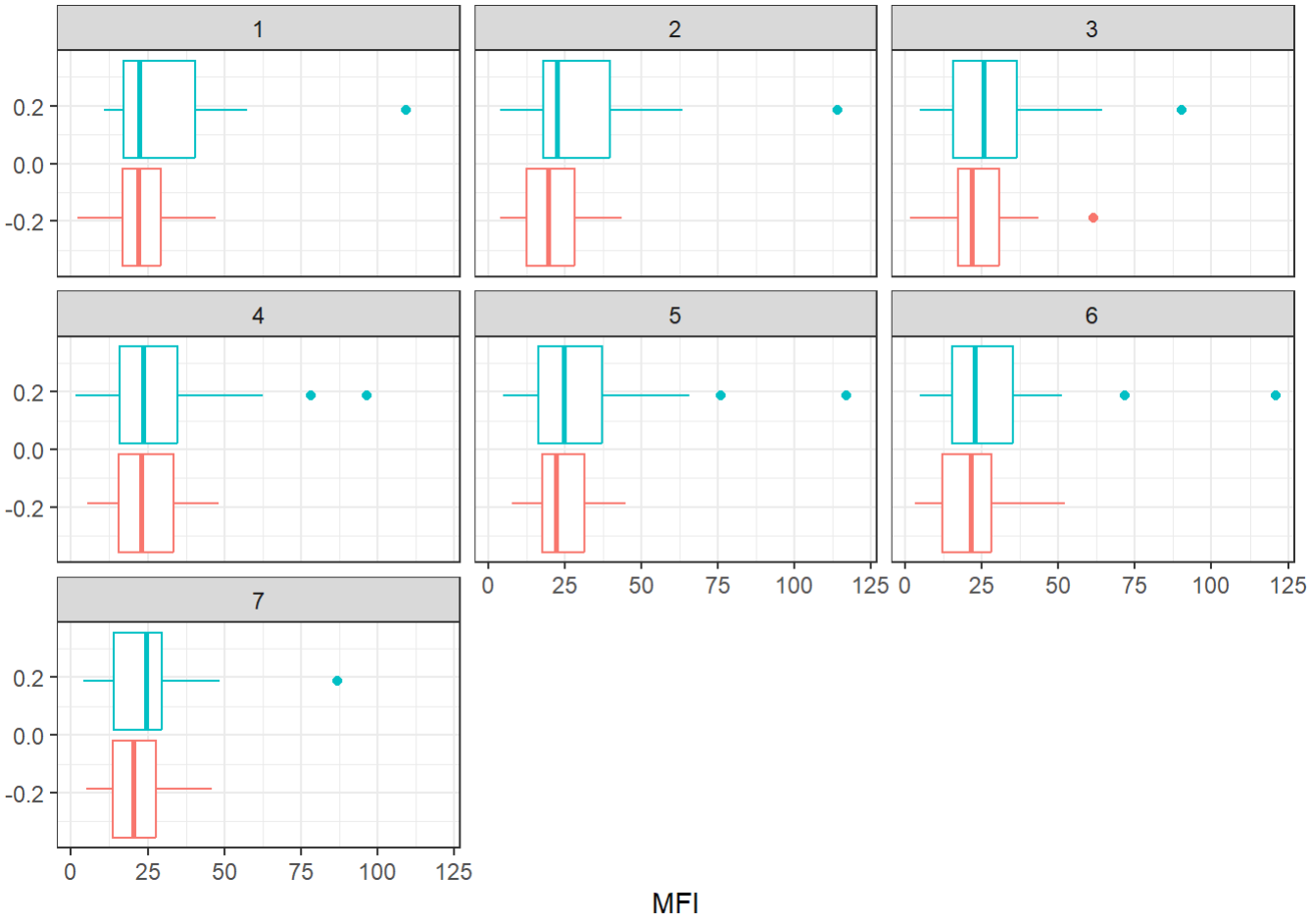
FHA is found on cell surfaces and coats the bacteria, helping it adhere to

> Q15 filter and plot two antigens for analysis

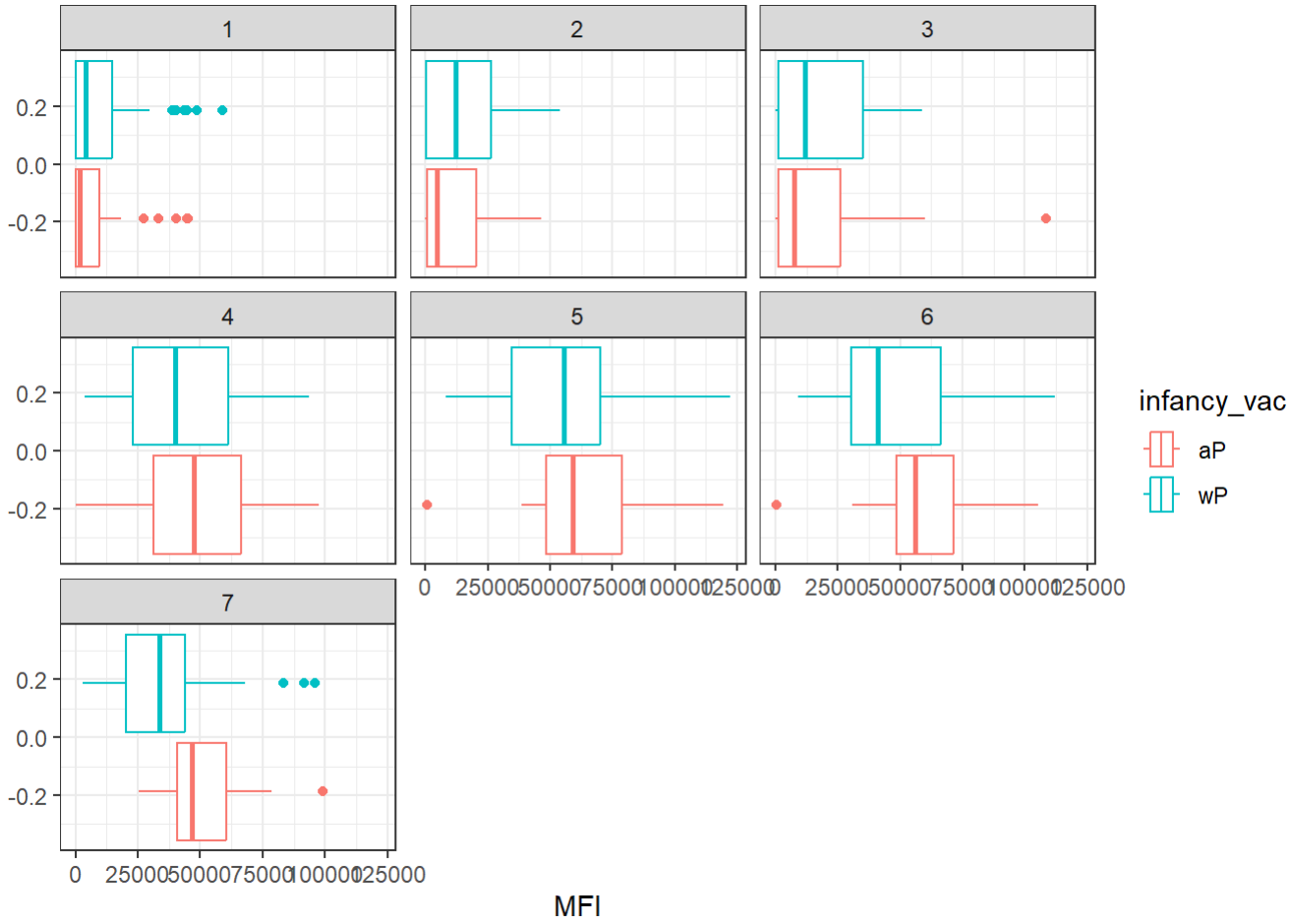
```
ggplot(ig1) +
  aes(MFI, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  theme_bw()
```




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




```
filter(ig1, antigen== 'FIM2/3') %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = T) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



> Q16 What do you notice about the time course for the antigens

The MFI for Measeles peaks around weeks 3-4 while the FIMM23 peaks in weeks 5-6.

> Q17 Do you see any clear differences in aP vs wP responses?

The differences between aP and wP response is more pronounced for the FIMM23 plots. This differences is especially noticable after the peak week 5, where the aP individuals have a higher MFI.

#obtaining CMI-PB RNAseq data

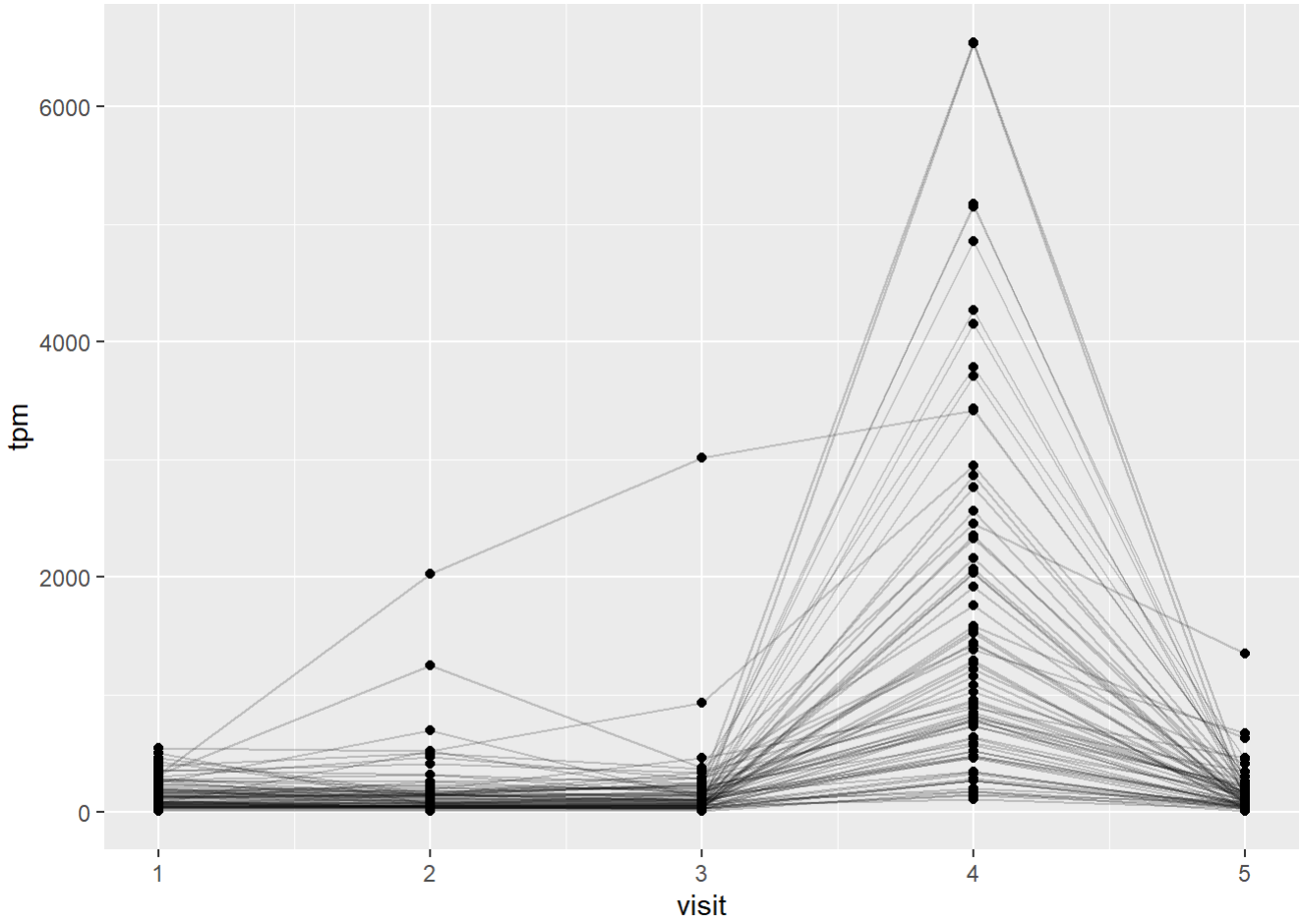
> Q18 Make a plot of time course of gene expression for IGHG1

```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.7"

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

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

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

```
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?

It is at it's maximum during the fourth visit. It does match the antibody titer data which has a similar peak for many igG1 antibodies around the fourth visit.