

Class19

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

```
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)
  )

```

We can now plot the number of reported pertussis cases per year in the U.S

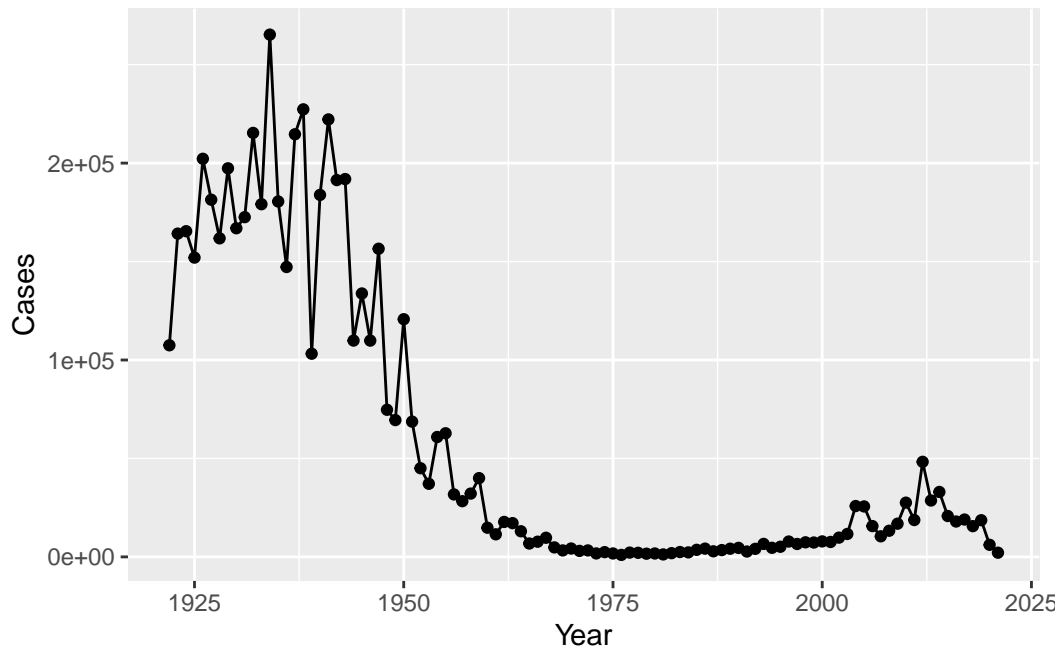
```
library(ggplot2)
```

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.

```

ggplot(cdc) +
  aes(Year, Cases) +
  geom_point() +
  geom_line()

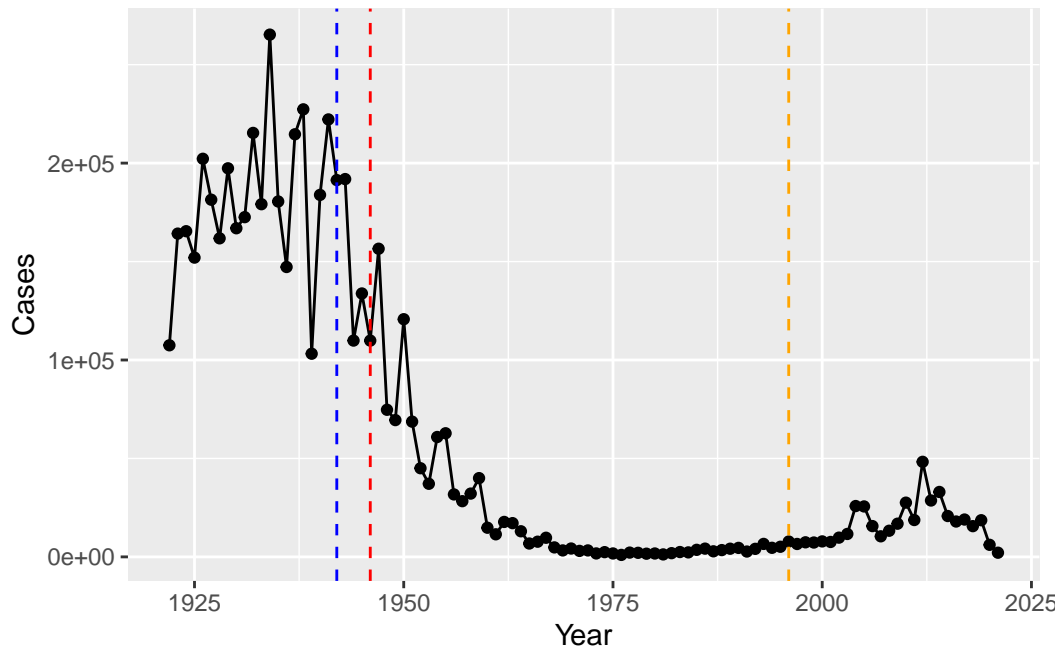
```



The first big “whole-cell” vaccine program started in 1942.

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. What do you notice?

```
ggplot(cdc) +
  aes(Year, Cases) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 1942, color = "blue", linetype = 2) +
  geom_vline(xintercept = 1946, color = "red", linetype = 2) +
  geom_vline(xintercept = 1996, color = "orange", linetype = 2)
```



Decrease in number of cases after introduction of wP vaccine. Slight increase in number of cases after introduction of 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 an increase in number of cases after introduction of aP vaccine. aP vaccine might be less effective than wP vaccine.

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

One of the main hypothesis for the increasing cases numbers is waning vaccine efficiency 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.

```
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?

```
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. Using this approach 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
```

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

```
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
```

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

```
round( summary( time_length( ap$age ) ) )
```

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

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

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

Yes. They are significantly different. There is about 9 year difference in mean age.

Q8. Determine the age of all individuals at time of boost?

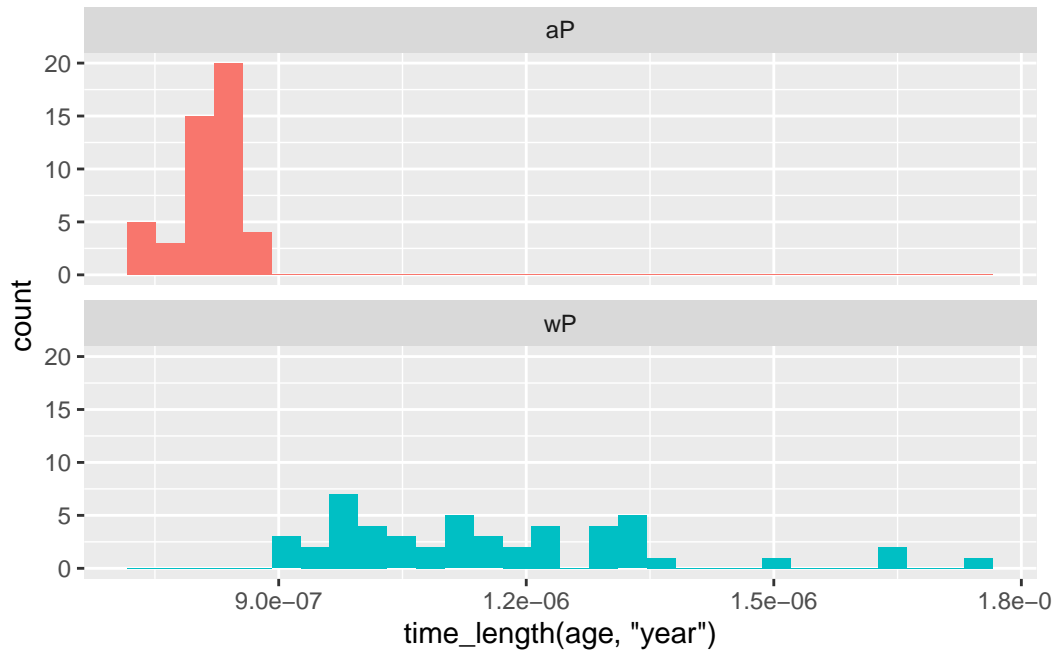
```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

```
[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`.



Now let's read some more database tables from OMI-PB:

```
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

I want to “join” (aka “merge”) the `subject` and `specimen` tables together. I will 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:

```
suppressMessages(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	37.42642	1
2	1986-01-01	2016-09-12	2020_dataset	37.42642	2
3	1986-01-01	2016-09-12	2020_dataset	37.42642	3
4	1986-01-01	2016-09-12	2020_dataset	37.42642	4
5	1986-01-01	2016-09-12	2020_dataset	37.42642	5
6	1986-01-01	2016-09-12	2020_dataset	37.42642	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 meta with titer data so we can further analyze this data in terms of time of visit aP/wP, male/female etc.

```

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

```

abdata <- inner_join(titer, meta)

```

Joining with `by = join_by(specimen_id)`

```

dim(abdata)

```

```

[1] 32675    21

```

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

```

So much smaller than the other visits because the project is still ongoing and we have not got that data for all individuals yet.

Examine IgG1 Ab titer levels

```

ig1 <- filter(abdata, 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 37.42642                      -3                      0
2 37.42642                      -3                      0
3 37.42642                      -3                      0
4 37.42642                      -3                      0
5 37.42642                      -3                      0
6 37.42642                      -3                      0
  specimen_type visit
1      Blood      1
2      Blood      1
3      Blood      1
4      Blood      1
5      Blood      1
6      Blood      1

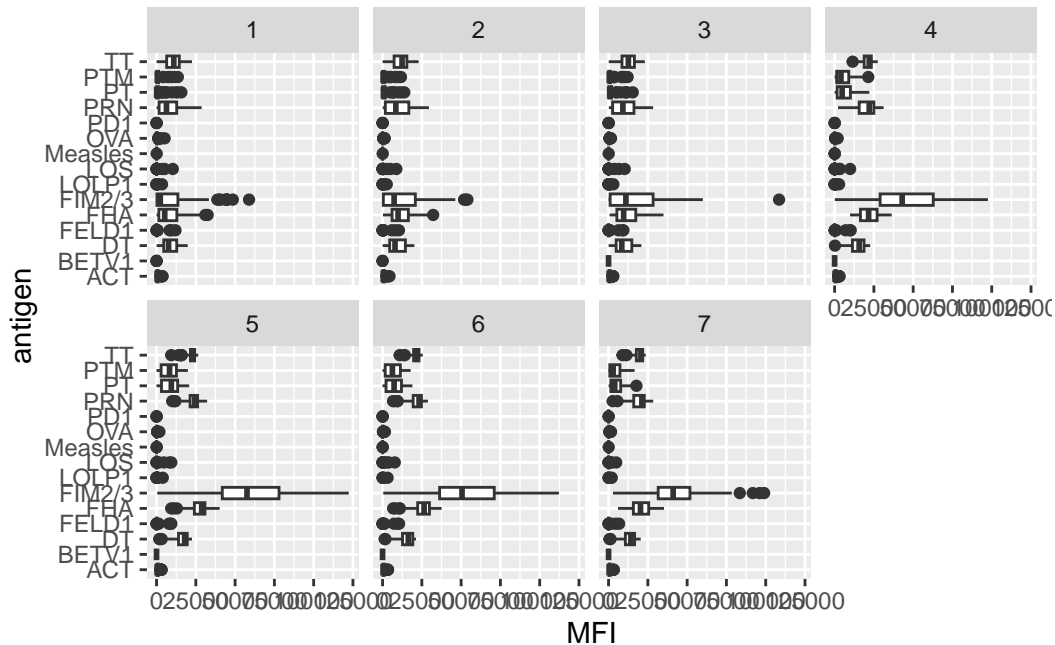
```

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?

FIM2/3 antigen shows difference in level of IgG1 antibody titers recognizing them over time. This is “Fimbrial protein” that makes the bacteria pilus and is involved in cell adhesion.

PT Pertussis toxin

FHA Filamentous hemagglutinin

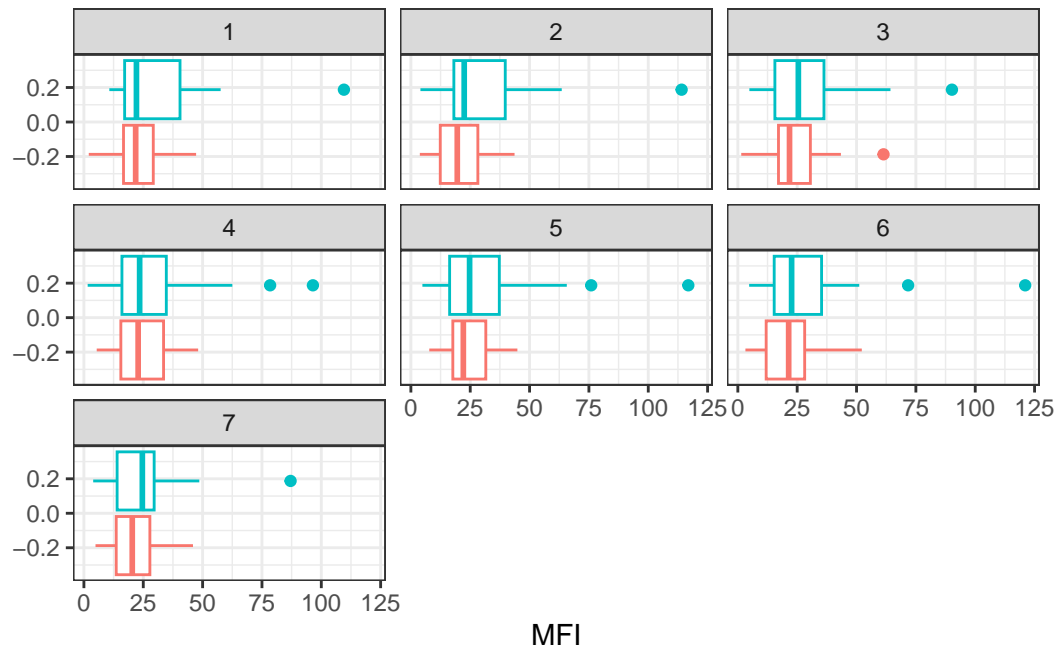
These antigens are expressed to be on the surface of the bacteria. Therefore IgG1 antibody can bind to it easily.

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

Measles antigen levels per visit (aP red; wP teal)

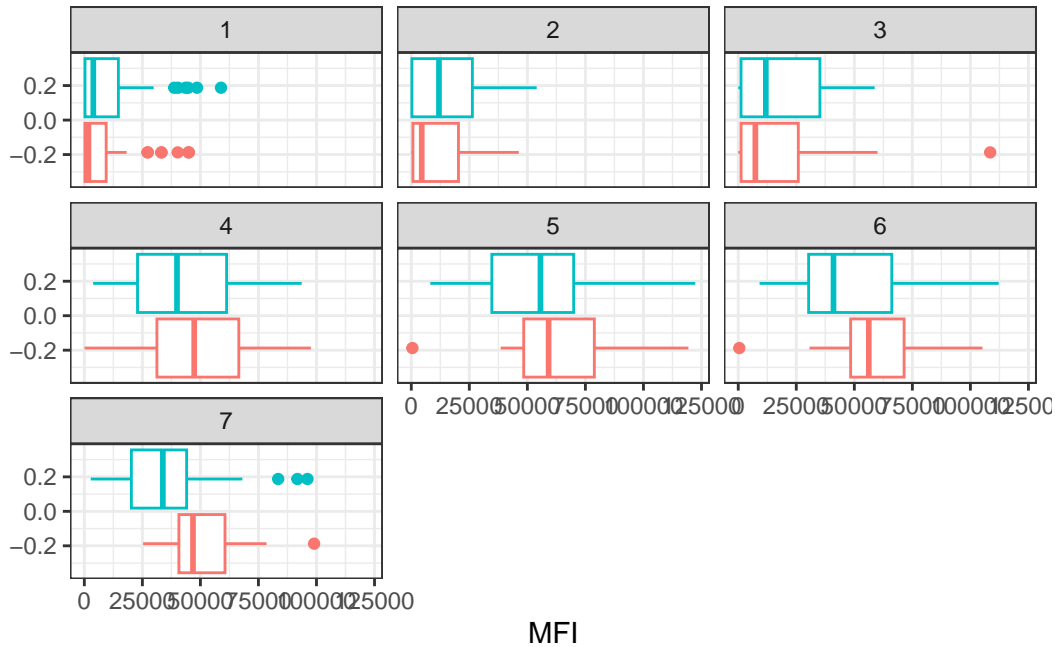
```
filter(ig1, antigen=="Measles") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
```

```
theme_bw()
```



FIM2/3 antigen levels per visit (aP red; wP teal)

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



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

For Measles, the median MFI seems does not really change per visit and there aren't much difference between aP and wP vaccine.

For FIM2/3, the median MFI increases per visit until the 7th visit where it drops. It seems the wP vaccine results in quicker reactions. It increases and drops more substantially than the MFI from aP vaccine.

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

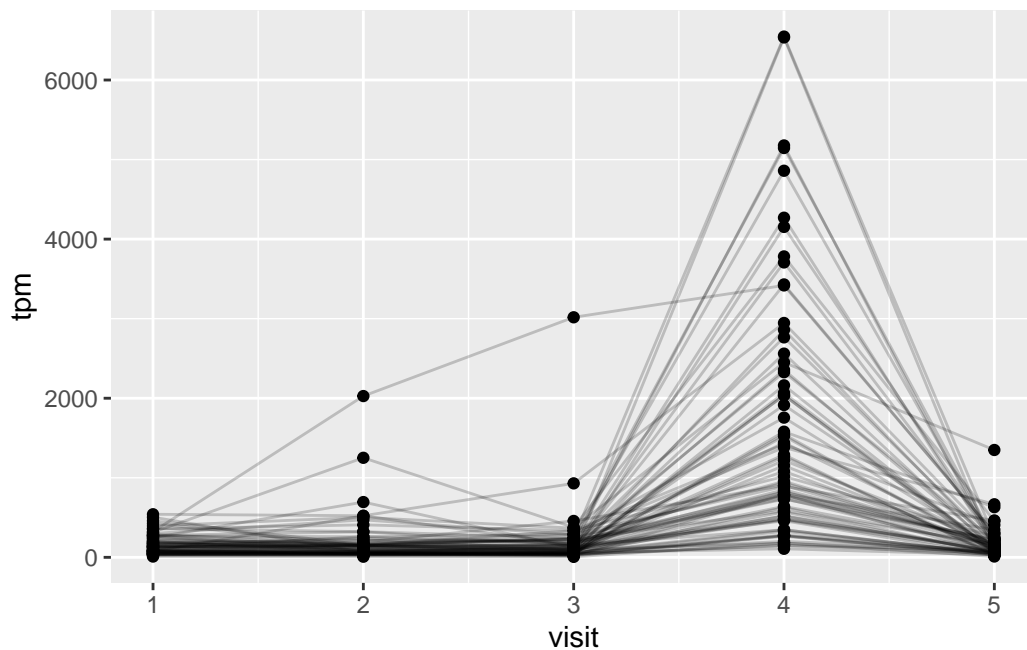
There is not a clear difference from the first few visits, but there is a clear difference in the 7th visit. Perhaps more data from visit 8 will make it more clear.

Q18. Make a plot of the time course of gene expression for IGHG1 gene (i.e. a plot of visit vs. tpm)

```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSOG00000211896."
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 (i.e. when is it at it's maximum level)?

It takes time for it to be expressed at a maximum level. Visit 4 had the highest level of expression.

Q20. Does this pattern in time match the trend of antibody titer data? If not, why not?

Yes it does because you see the MFI for the FIM2/3 level rise per visit and drop after some time.