

Pertussis

Pertussis, or whooping cough, is a highly contagious lung infection caused by a bacteria *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)
)
```

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.

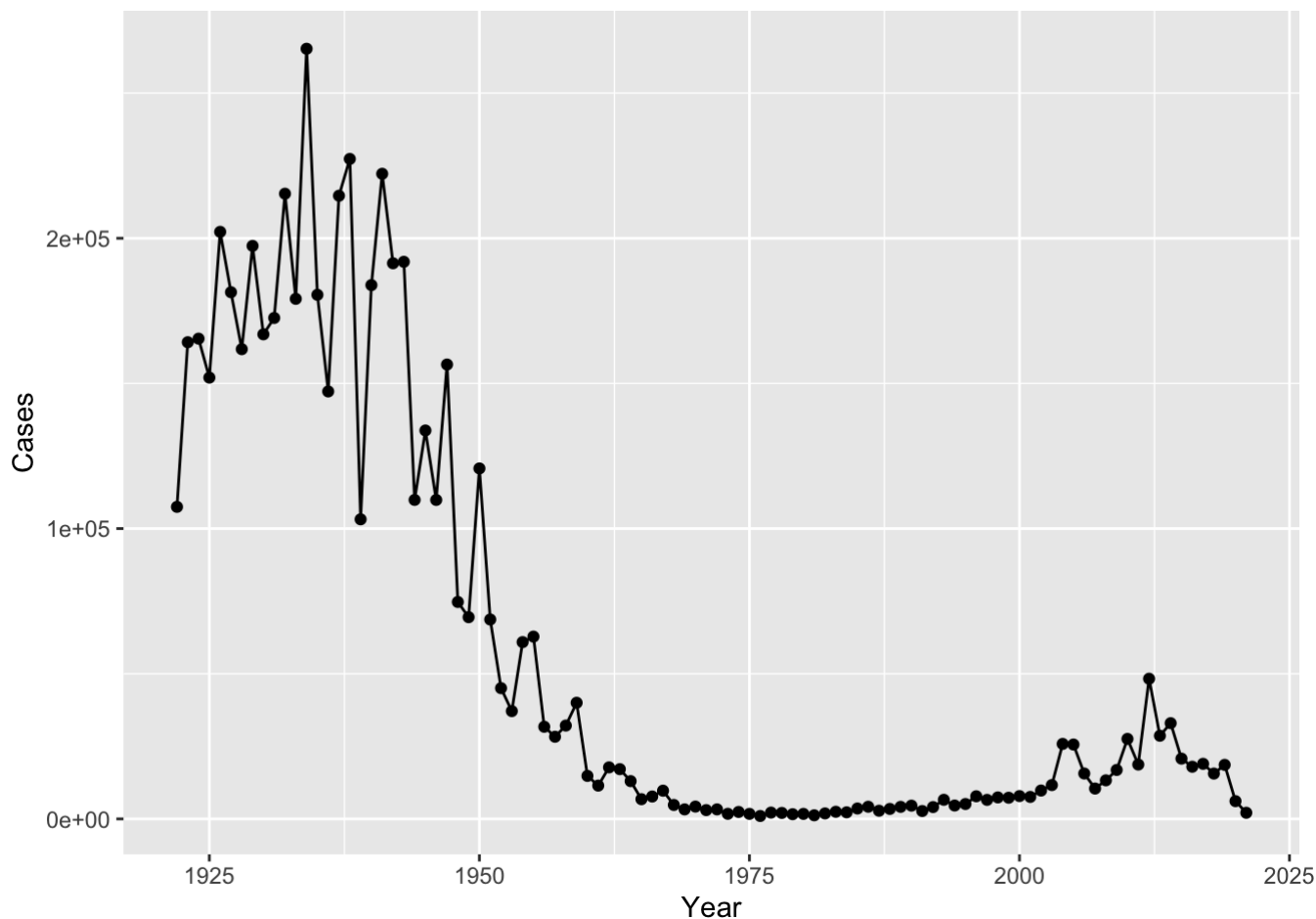
```
library(ggplot2)

ggplot(cdc) +
```

```

aes(Year, Cases) +
geom_point() +
geom_line() +
labs()

```



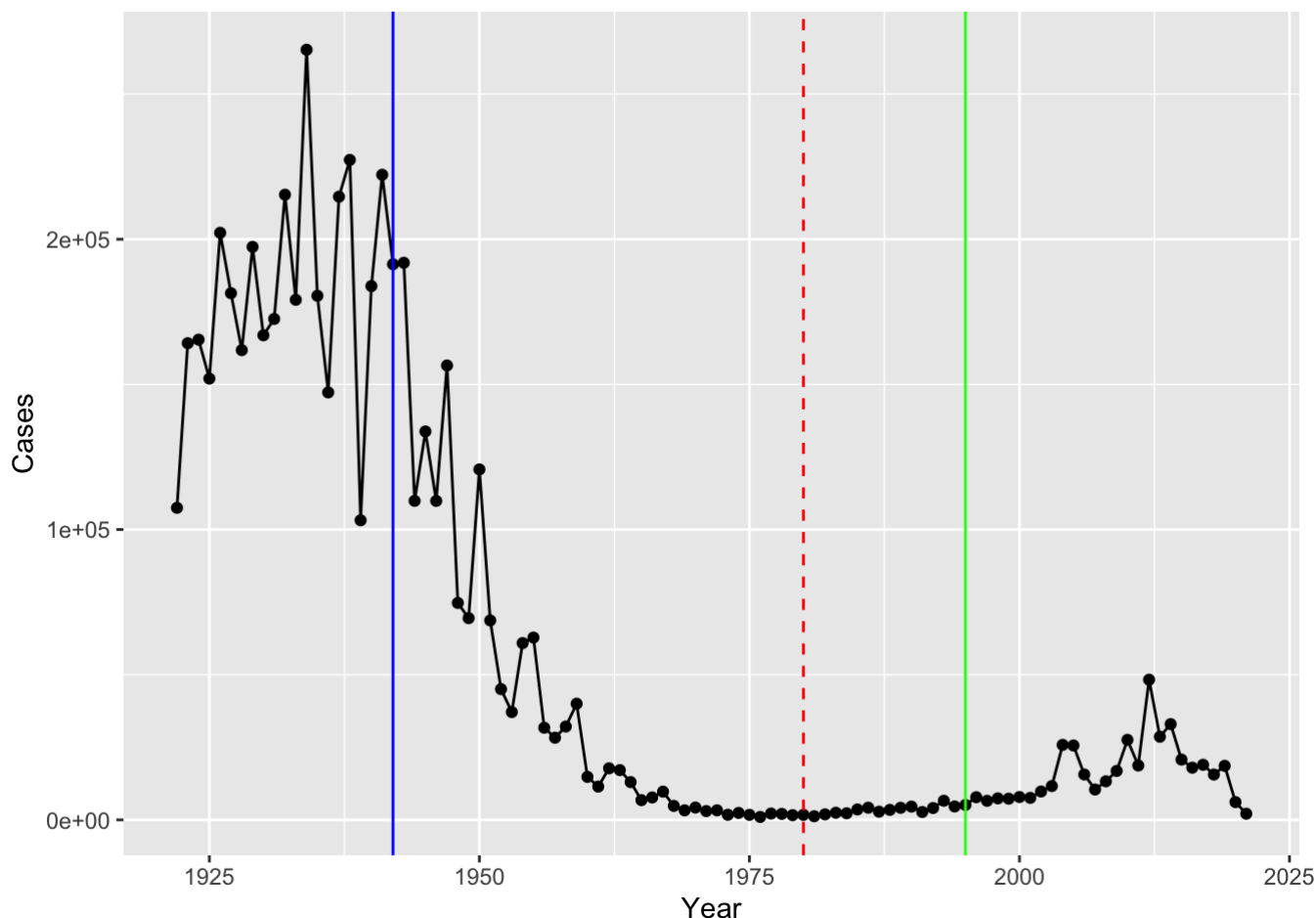
Q2. Using the ggplot `geom_vline()` function add lines to your previous plot for the 1946 introduction of the wP vaccine and the 1996 switch to aP vaccine (see example in the hint below). What do you notice?

```

library(ggplot2)

ggplot(cdc) +
  aes(Year, Cases) +
  geom_point() +
  geom_line() +
  labs() +
  geom_vline(xintercept = 1942, color="blue") +
  geom_vline(xintercept = 1980, color="red", linetype= 2) +
  geom_vline(xintercept = 1995, color="green")

```



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

It is clear from the CDC data that pertussis cases are once again increasing. For example, we can see that in 2012 the CDC reported 48,277 cases of pertussis in the United States. This is the largest number of cases reported since 1955, when 62,786 cases were reported. The pertussis field has several hypotheses for the resurgence of pertussis including (in no particular order): 1) more sensitive PCR-based testing, 2) vaccination hesitancy 3) bacterial evolution (escape from vaccine immunity), 4) waning of immunity in adolescents originally primed as infants with the newer aP vaccine as compared to the older wP vaccine.

```
# Allows us to read, write and process JSON 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	White
2	2	wP	Female Not Hispanic or Latino	White	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
```

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

```
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

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)
```

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(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	specimen_id	
1	1986-01-01	2016-09-12	2020_dataset	1	
2	1986-01-01	2016-09-12	2020_dataset	2	
3	1986-01-01	2016-09-12	2020_dataset	3	
4	1986-01-01	2016-09-12	2020_dataset	4	
5	1986-01-01	2016-09-12	2020_dataset	5	
6	1986-01-01	2016-09-12	2020_dataset	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		

```
library(lubridate)
```

Attaching package: 'lubridate'

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

date, intersect, setdiff, union

```
today()
```

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

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

Time difference of 8557 days

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

```
[1] 23.42779
```

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?

```
# Assuming 'subject$year_of_birth' contains birth year of subjects in "yyyy-mm-dd" format

# Use today's date to calculate age in days
subject$age <- today() - ymd(subject$year_of_birth)

# aP
ap <- subject %>% filter(infancy_vac == "aP")
round( mean( time_length( ap$age, "years" ) ), 1 )
```

```
[1] 25.7
```

```
# wP
wp <- subject %>% filter(infancy_vac == "wP")
round( summary( time_length( wp$age, "years" ) ), 1 )
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
28.4	32.4	35.4	36.6	40.4	55.4

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

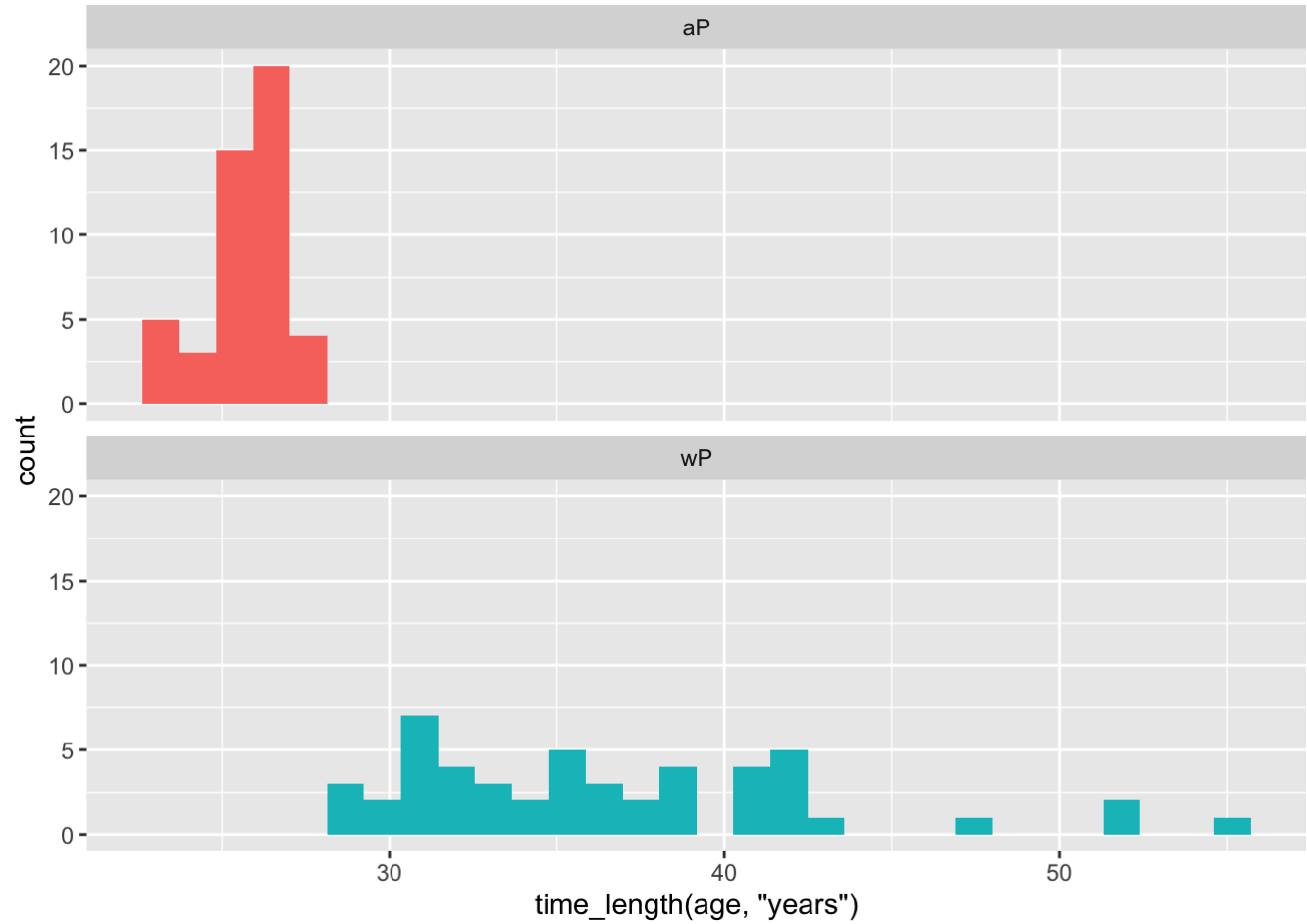
```
# To determine whether they are significantly different, we can use a t-test
subject$age_at_boost <- time_length(ymd(subject$date_of_boost) - ymd(subject$year_of_birt
head(subject$age_at_boost)
```

```
[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?

```
# Q9
ggplot(subject) +
  aes(x=time_length(age, "years"),
      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`.



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.

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

head(ab)
```

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(meta, ab)
```

Joining with `by = join_by(specimen_id)`

```
dim(abdata)
```

```
[1] 32675    20
```

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

```

```
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

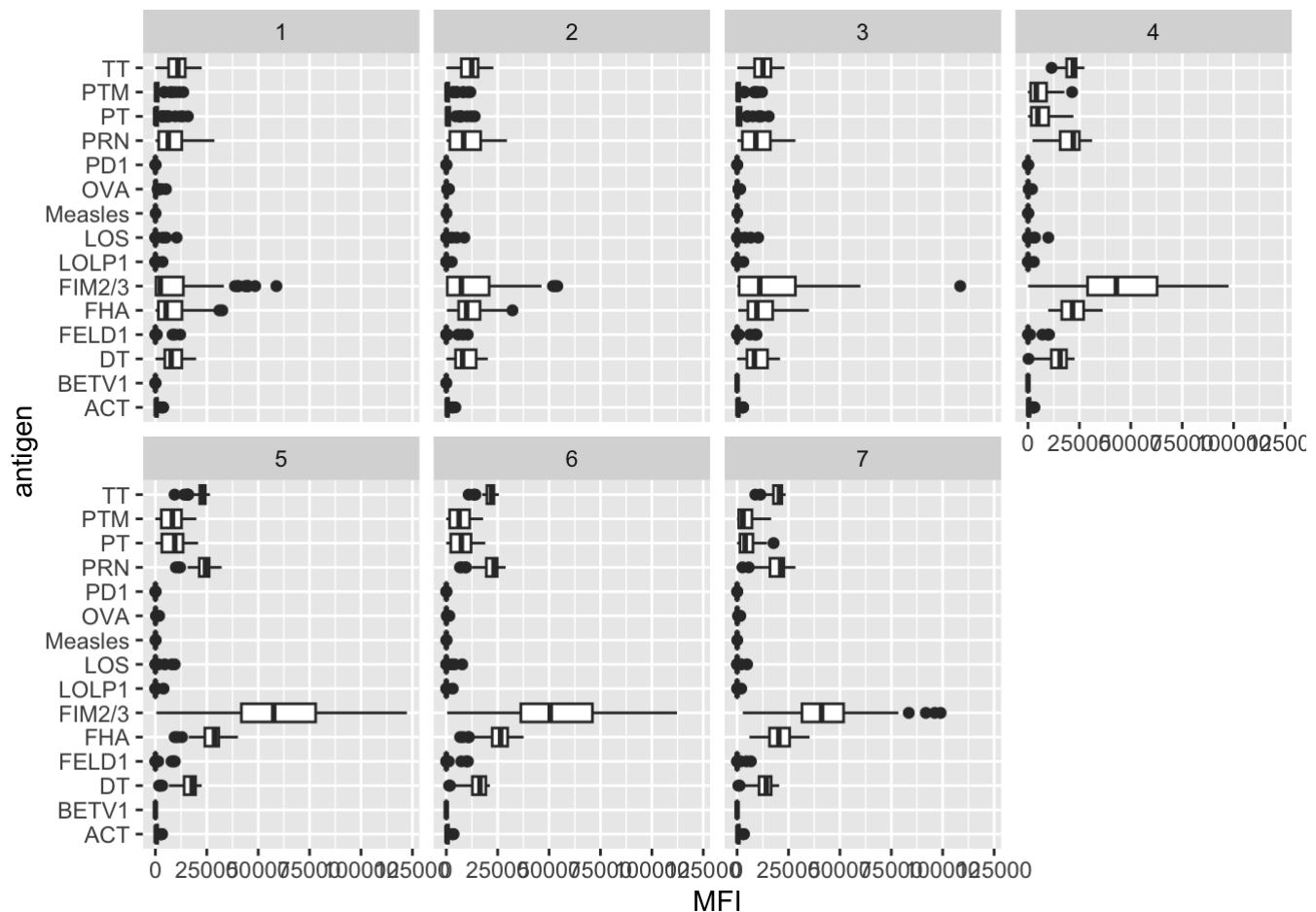
	year_of_birth	date_of_boost	dataset	specimen_id
1	1986-01-01	2016-09-12	2020_dataset	1
2	1986-01-01	2016-09-12	2020_dataset	1
3	1986-01-01	2016-09-12	2020_dataset	1
4	1986-01-01	2016-09-12	2020_dataset	1
5	1986-01-01	2016-09-12	2020_dataset	1
6	1986-01-01	2016-09-12	2020_dataset	1

	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:

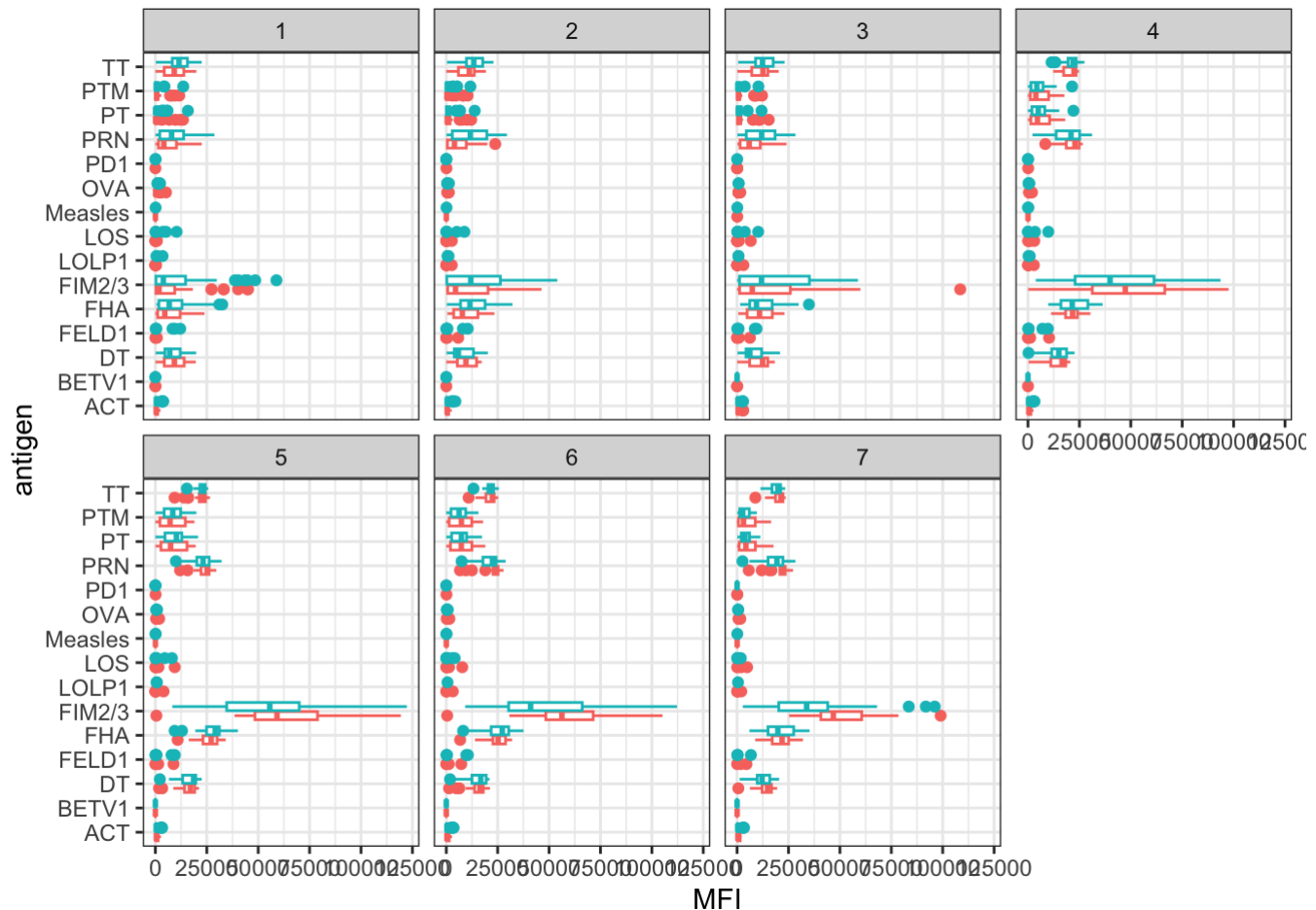
```
ggplot(ig1) +
  aes(MFI, antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow=2)
```



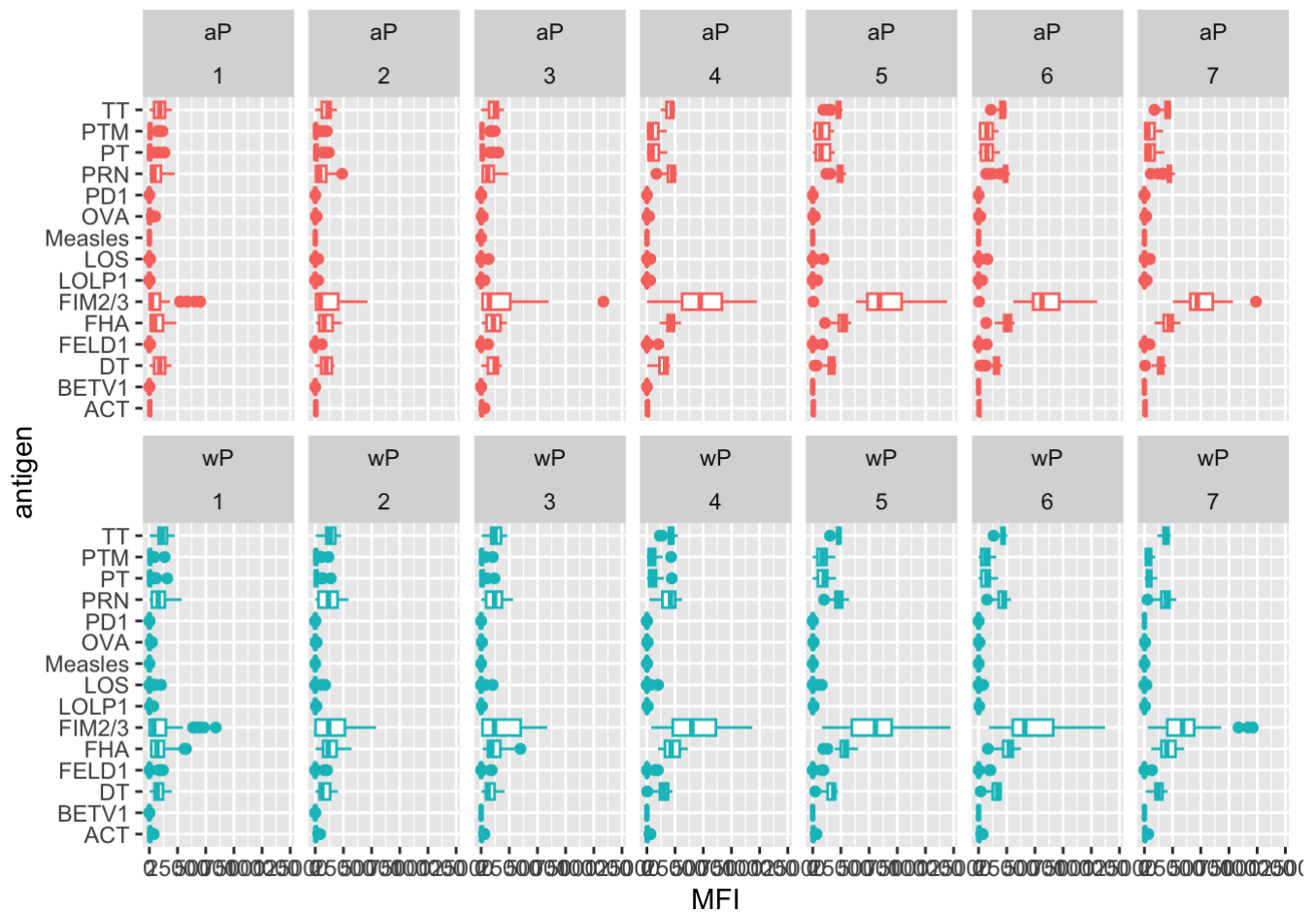
Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

Fim2/3

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

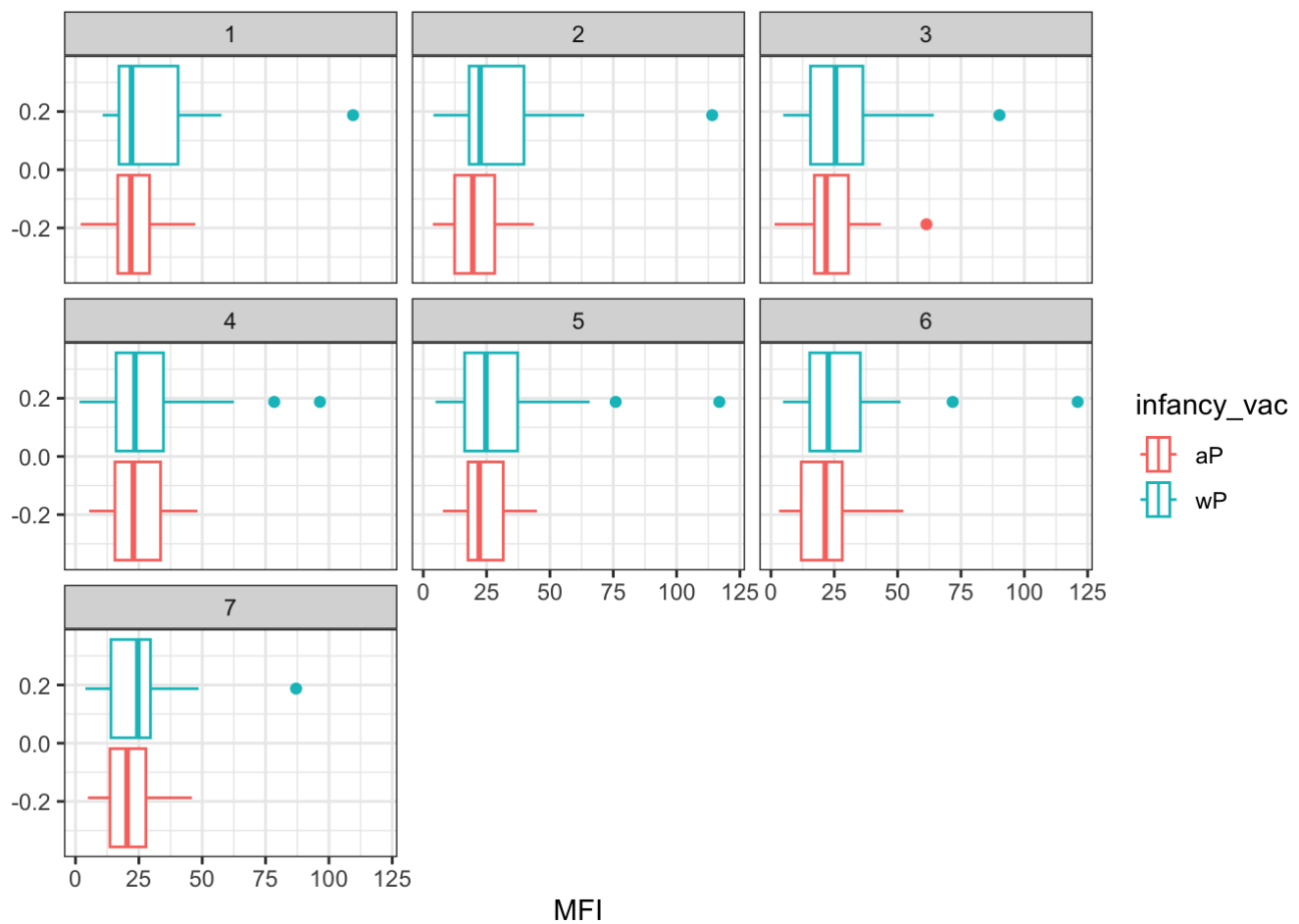


```
ggplot(ig1) +
  aes(MFI, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```

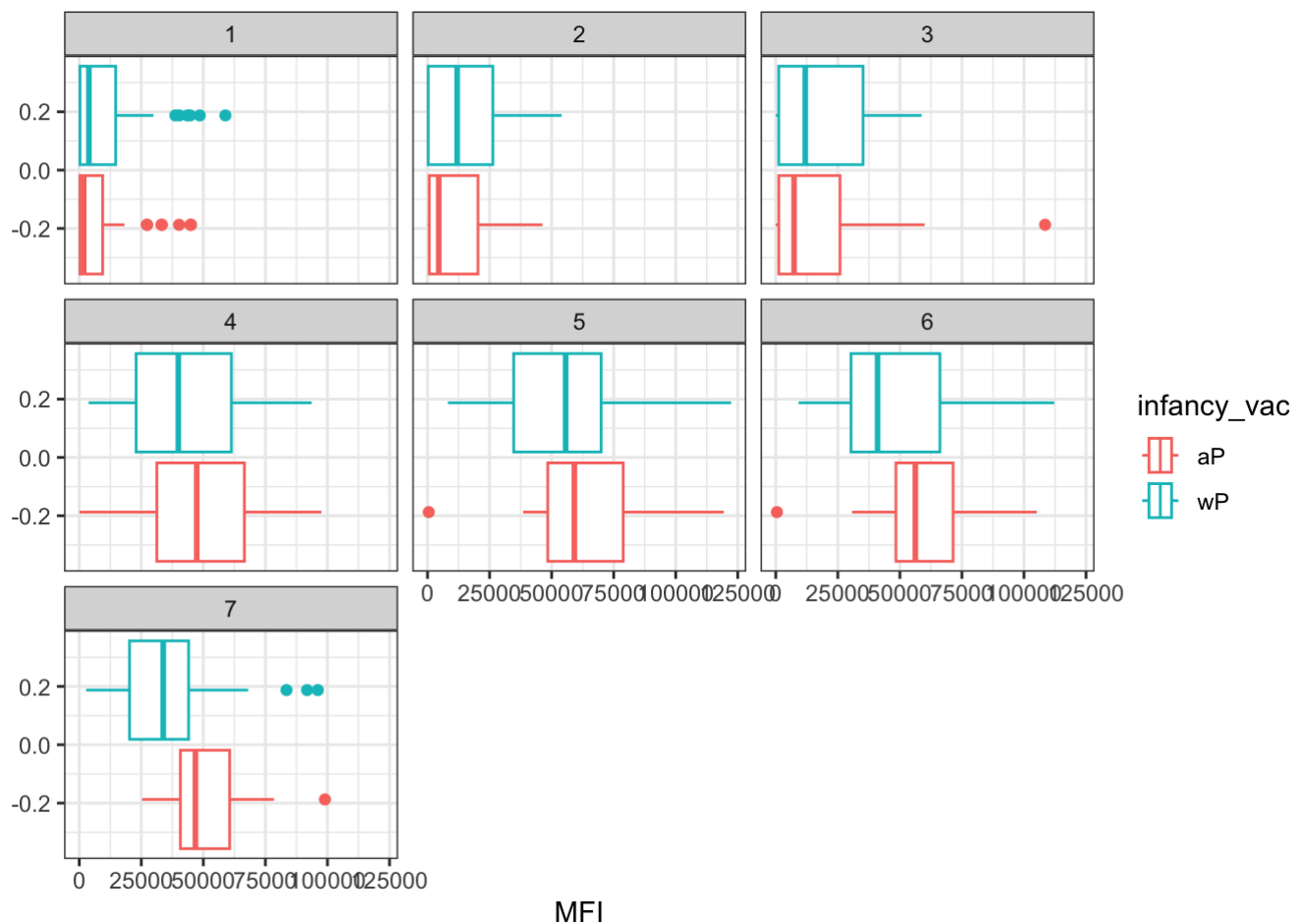


Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("Measles", that is not in our vaccines) and a clear antigen of interest ("FIM2/3", extra-cellular fimbriae proteins from *B. pertussis* that participate in substrate attachment).

```
filter(ig1, antigen=="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 clearly rise over time and far exceed those of Measles. They also appear to peak at visit 5 and then decline. This trend appears similar for for wP and aP subjects.

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

It appears that aP clearly rise and are more elevated compared to wP over time.

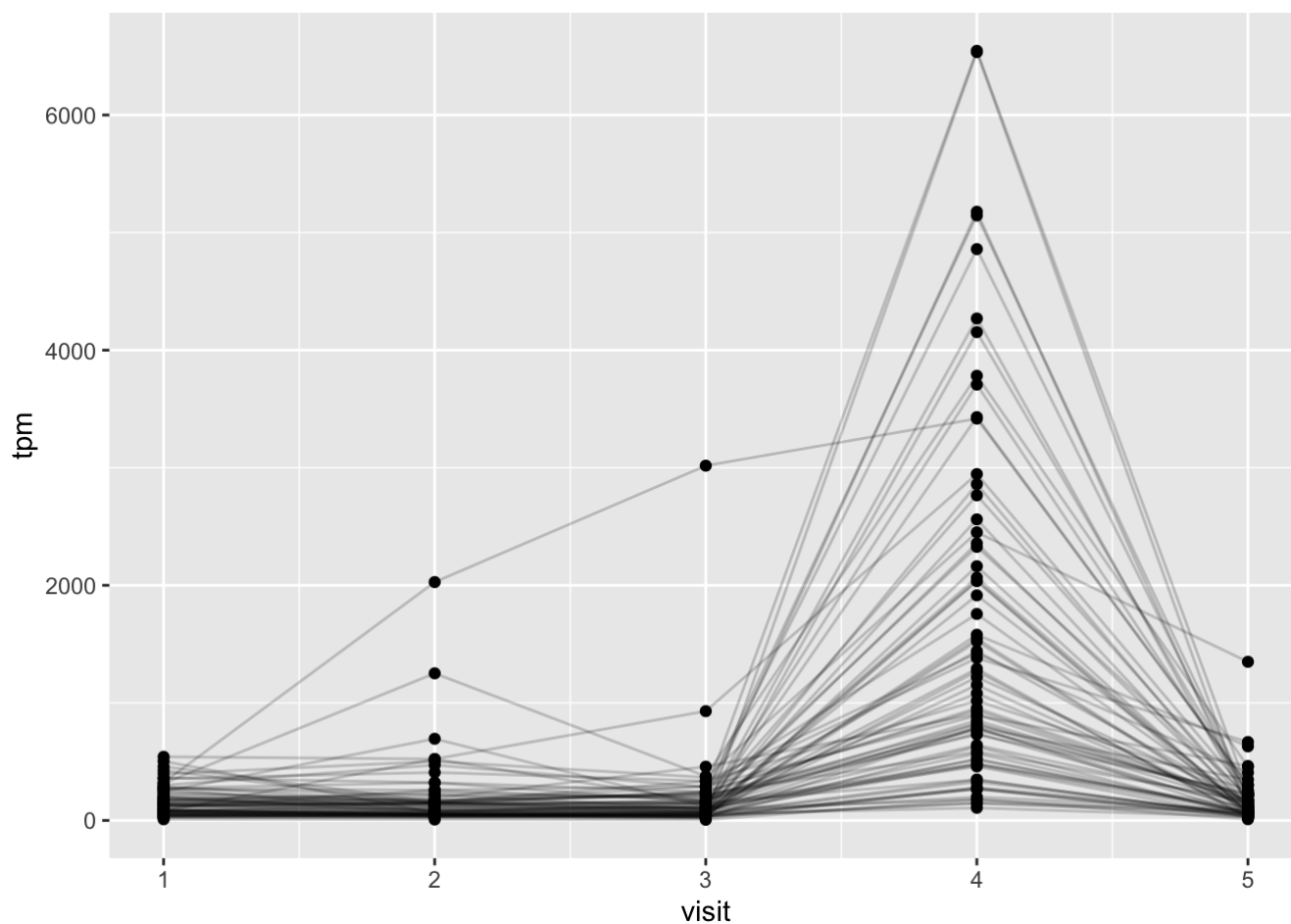
```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENS00000211896"
rna <- read_json(url, simplifyVector = TRUE)
```

```
#meta <- inner_join(specimen, subject)
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 (i.e. when is it at it's maximum level)?

It's most highly expressed at visit 4.

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

Cells make antibodies, which are long lived