

lab15_pertussis

Longmei Zhang A17012012

Background

Pertussis (more commonly known as whooping cough) is a highly contagious respiratory disease caused by the bacterium *Bordetella pertussis*. It can infect people of all ages, causing high mortality rates in infants.

##Investigating pertussis cases by year

We can view this data on the CDC website here: [CDC data](#)

Import the pertussis data from the website. We will use the **datapasta** R package to scrape this data into R. Use Addins **paste as data.frame**

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,2022L),

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,
6124,2116,3044)

)

```

```
head(cdc)
```

```

  year  cases
1 1922 107473
2 1923 164191
3 1924 165418
4 1925 152003
5 1926 202210
6 1927 181411

```

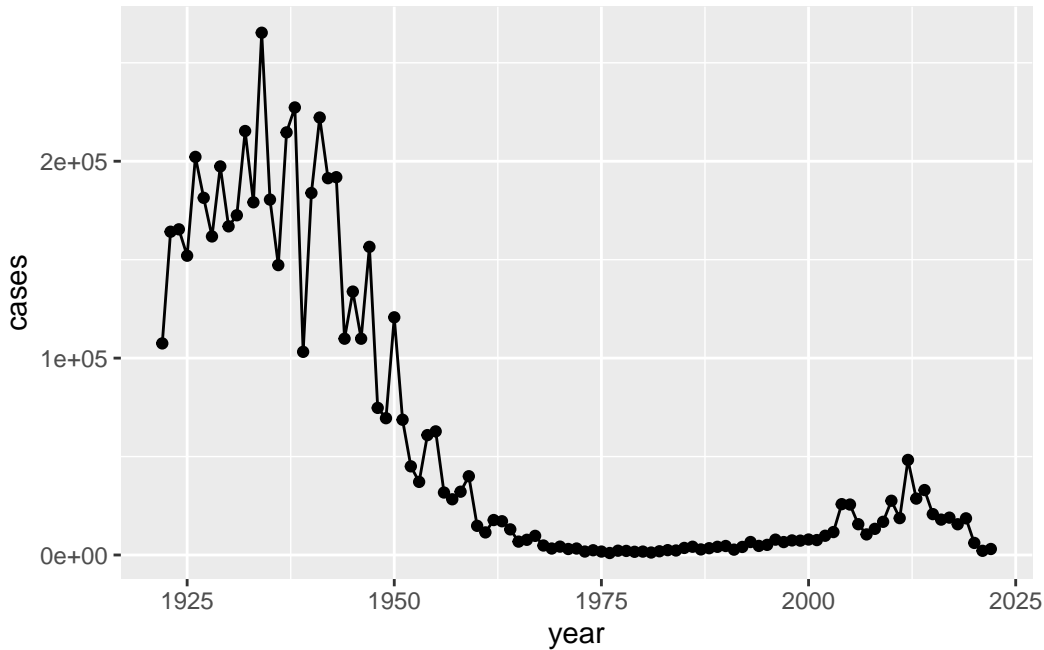
```
library(ggplot2)
```

```

cdc_plot <- ggplot(cdc) +
  aes(year, cases) +
  geom_point() +
  geom_line()

```

```
cdc_plot
```



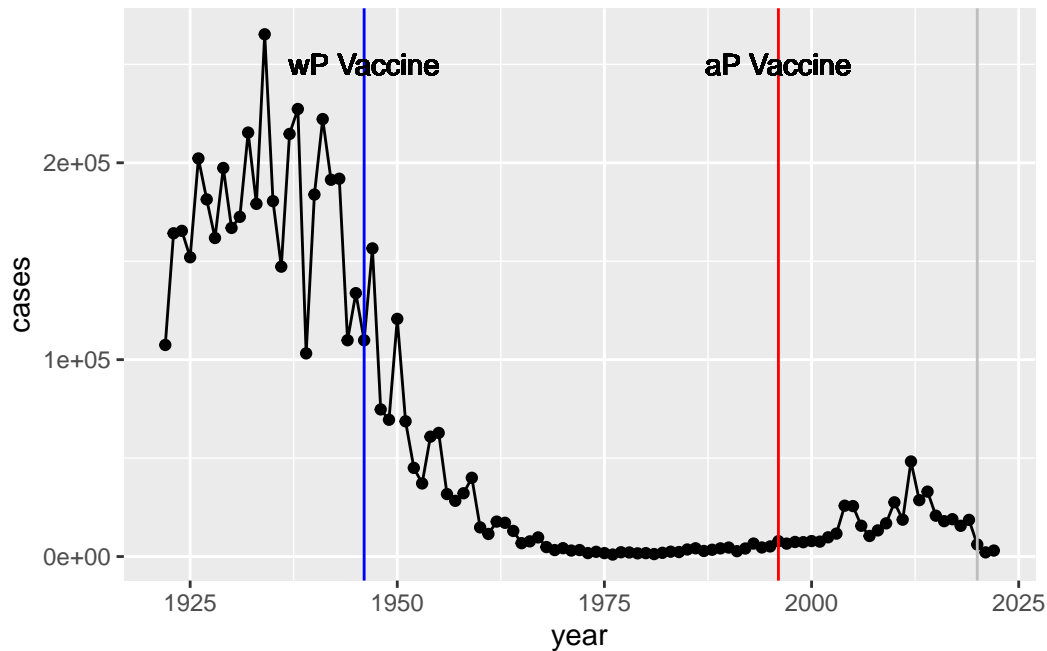
Two Vaccines: wP & aP

Adding some landmark developments as annotations to our plot.

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?

the number of cases significantly decreased after the introduction of wP vaccine in 1946. We went from ~200,000 cases per year before wP vaccine to ~1000 cases per year after wP vaccine. The plot shows an increasing trend after the switch to aP vaccine after 1996. We see a big increase in 2004 to 26,000 cases.

```
cdc_plot +  
  geom_vline(xintercept = 1946, col = "blue") +  
  geom_vline(xintercept = 1996, col = "red") +  
  geom_vline(xintercept = 2020, col = "grey") +  
  geom_text(x = 1996, y = 250000, label = "aP Vaccine") +  
  geom_text(x = 1946, y = 250000, label = "wP Vaccine")
```



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

The number of cases increased after the switch to aP vaccine in 1996. The aP vaccine might be less effective comparing to wP vaccines. The aP vaccine does not provide long term immunity against the bacteria. The Covid pandemic might also contributed to the increase in pertussis cases.

There is a ~10 year lag from switching to aP vaccine to increase in Pertussis cases. This holds true for different countries.

Key question: why does the aP vaccine induced immunity wane faster than that of the wP vaccine?

CMI-PB data

The CMI-PB (computational Models of Immunity Pertussis Boost) makes available lots of data about the immune response to Pertussis vaccination.

Critically, it tracks wP and aP individuals over time to see how their immune response changes.

CMI-PB make all their data freely available via JSON for tables from their database

Importing the first table using package `json`

```
library(jsonlite)
```

Warning: package 'jsonlite' was built under R version 4.3.3

```
subject <- read_json("https://www.cmi-pb.org/api/v5/subject", simplifyVector = T)
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

Q. How many subjects are there in this dataset?

```
nrow(subject)
```

```
[1] 172
```

Q4. How many aP or wP?

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

```
aP wP
87 85
```

Q5. How many male or female

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

```
Female    Male
   112     60
```

Q6. Breakdown by biological_sex and race. e.g. how many black female subjects etc.

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

	Female	Male
American Indian/Alaska Native	0	1
Asian	32	12
Black or African American	2	3
More Than One Race	15	4
Native Hawaiian or Other Pacific Islander	1	1
Unknown or Not Reported	14	7
White	48	32

Q. Does this study do a good job of representing US population

Strong bias toward white individuals. Other ethnicities, especially American Indian/Alaska Native, Black or African American, and Native Hawaiian or Other Pacific Islander are neglected.

Import the second table for specimen for more data

```
specimen <- read_json("https://www.cmi-pb.org/api/v5/specimen", simplifyVector=T)
head(specimen)
```

	specimen_id	subject_id	actual_day_relative_to_boost	
1	1	1	-3	
2	2	1	1	
3	3	1	3	
4	4	1	7	
5	5	1	11	
6	6	1	32	

	planned_day_relative_to_boost	specimen_type	visit
1	0	Blood	1

2	1	Blood	2
3	3	Blood	3
4	7	Blood	4
5	14	Blood	5
6	30	Blood	6

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?

wp individuals seem to be significantly older than the ap individuals. The mean age of ap individual is 27, while the mean age of wp individual is 36.

```
library(lubridate)
```

Attaching package: 'lubridate'

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

date, intersect, setdiff, union

```
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

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

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
22	26	27	27	28	34

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

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
22	32	34	36	39	57

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

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

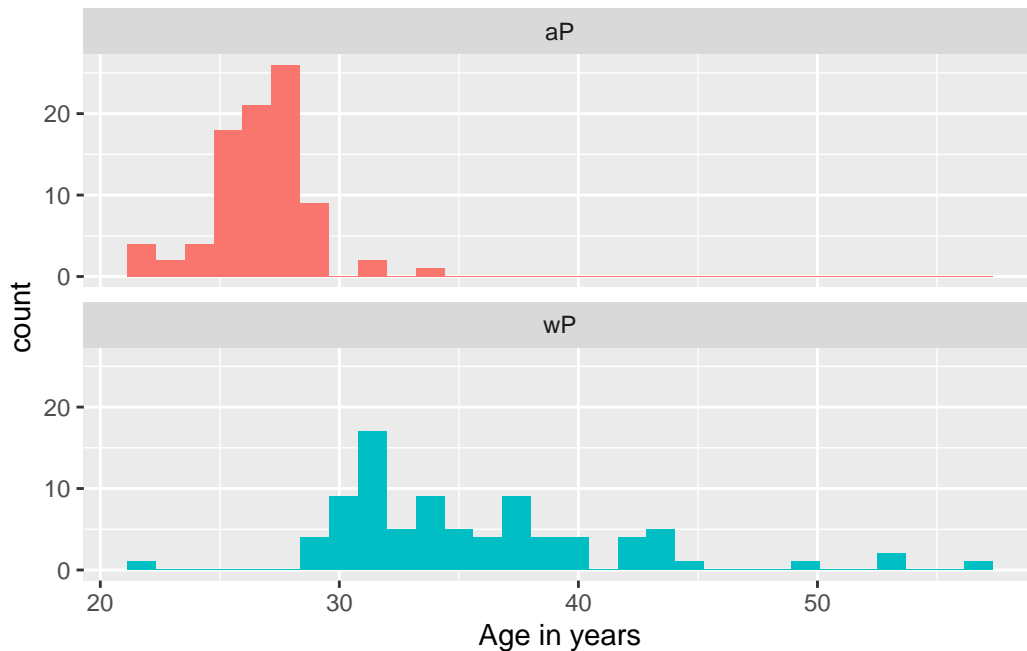
```
[1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481
```

Q9. With the help of a faceted boxplot or histogram (see below), do you think these two groups are significantly different?

The age of individuals from 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`.



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:

Now we can join (merge) these two tables specimen and subject for a meta table with combined information.

```
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	age_at_boost	specimen_id
1	1986-01-01	2016-09-12	2020_dataset	14208 days	30.69678	1
2	1986-01-01	2016-09-12	2020_dataset	14208 days	30.69678	2

```

3 1986-01-01 2016-09-12 2020_dataset 14208 days 30.69678 3
4 1986-01-01 2016-09-12 2020_dataset 14208 days 30.69678 4
5 1986-01-01 2016-09-12 2020_dataset 14208 days 30.69678 5
6 1986-01-01 2016-09-12 2020_dataset 14208 days 30.69678 6
  actual_day_relative_to_boost planned_day_relative_to_boost specimen_type
1                -3                0      Blood
2                 1                1      Blood
3                 3                3      Blood
4                 7                7      Blood
5                11               14      Blood
6                32               30      Blood
  visit
1     1
2     2
3     3
4     4
5     5
6     6

```

```

titer <- read_json("https://www.cmi-pb.org/api/v5/plasma_ab_titer", simplifyVector = T)
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

```

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.

Join the previous meta table with the new antibody table to associate all the metadata about the individual and their race, biological sex, and infancy vaccination status together with Antibody levels

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

Joining with `by = join_by(specimen_id)`

```
head(abdata)
```

	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	subject_id	infancy_vac	biological_sex
1	UG/ML	2.096133	1	wP	Female
2	IU/ML	29.170000	1	wP	Female
3	IU/ML	0.530000	1	wP	Female
4	IU/ML	6.205949	1	wP	Female
5	IU/ML	4.679535	1	wP	Female
6	IU/ML	2.816431	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	age_at_boost	actual_day_relative_to_boost
1	14208 days	30.69678	-3
2	14208 days	30.69678	-3
3	14208 days	30.69678	-3
4	14208 days	30.69678	-3
5	14208 days	30.69678	-3
6	14208 days	30.69678	-3

	planned_day_relative_to_boost	specimen_type	visit
1	0	Blood	1
2	0	Blood	1
3	0	Blood	1
4	0	Blood	1
5	0	Blood	1
6	0	Blood	1

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  5389 10117 10124 10124 10124
```

How many antigens

```
table(abdata$antigen)
```

```
  ACT  BETV1    DT  FELD1    FHA  FIM2/3  LOLP1    LOS Measles    OVA
1970   1970   4978   1970   5372   4978   1970   1970   1970   4978
  PD1    PRN    PT    PTM  Total    TT
1970   5372   5372   1970    788   4978
```

Q12. What are the different \$dataset values in abdata and what do you notice about the number of rows for the most “recent” dataset?

different \$dataset values are the dataset in different years that contain the data. The most recent dataset has 5670 rows.

```
table(abdata$dataset)
```

```
2020_dataset 2021_dataset 2022_dataset 2023_dataset
      31520         8085         7301         5670
```

Examine IgG Ab titer levels

Lets focus on IgG - one of the main antibody types respond to bacterial or viral infections

```
igg <- filter(abdata, isotype == "IgG")
head(igg)
```

	specimen_id	isotype	is_antigen_specific	antigen	MFI	MFI_normalised
1	1	IgG	TRUE	PT	68.56614	3.736992
2	1	IgG	TRUE	PRN	332.12718	2.602350
3	1	IgG	TRUE	FHA	1887.12263	34.050956
4	19	IgG	TRUE	PT	20.11607	1.096366
5	19	IgG	TRUE	PRN	976.67419	7.652635
6	19	IgG	TRUE	FHA	60.76626	1.096457

	unit	lower_limit_of_detection	subject_id	infancy_vac	biological_sex
1	IU/ML	0.530000	1	wP	Female
2	IU/ML	6.205949	1	wP	Female
3	IU/ML	4.679535	1	wP	Female
4	IU/ML	0.530000	3	wP	Female
5	IU/ML	6.205949	3	wP	Female
6	IU/ML	4.679535	3	wP	Female

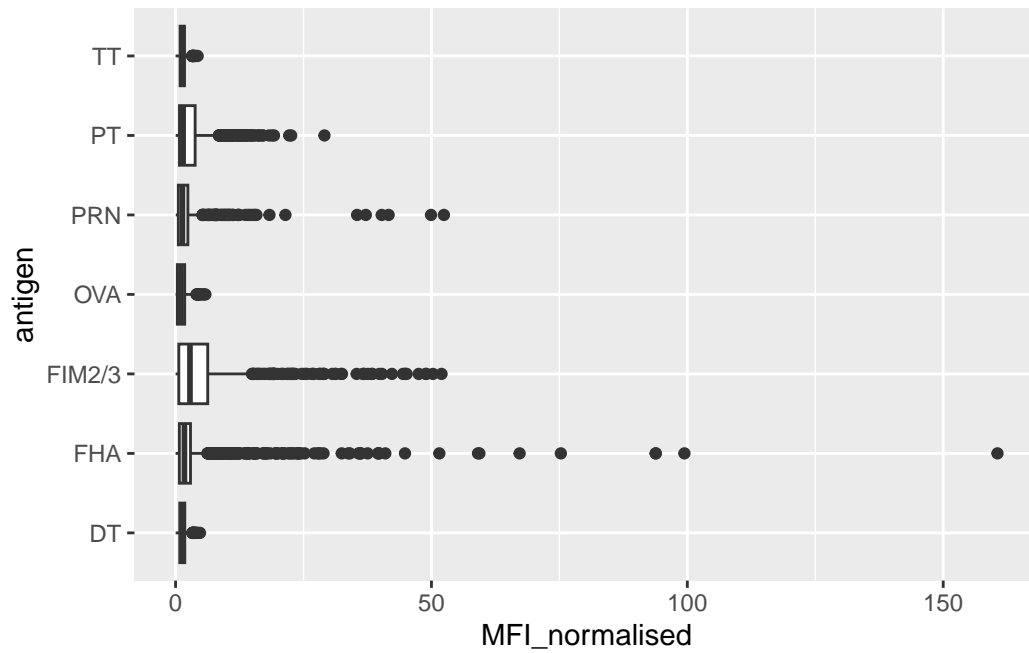
	ethnicity	race	year_of_birth	date_of_boost	dataset
1	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
2	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
3	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
4	Unknown	White	1983-01-01	2016-10-10	2020_dataset
5	Unknown	White	1983-01-01	2016-10-10	2020_dataset
6	Unknown	White	1983-01-01	2016-10-10	2020_dataset

	age	age_at_boost	actual_day_relative_to_boost
1	14208 days	30.69678	-3
2	14208 days	30.69678	-3
3	14208 days	30.69678	-3
4	15304 days	33.77413	-3
5	15304 days	33.77413	-3
6	15304 days	33.77413	-3

	planned_day_relative_to_boost	specimen_type	visit
1	0	Blood	1
2	0	Blood	1
3	0	Blood	1
4	0	Blood	1
5	0	Blood	1
6	0	Blood	1

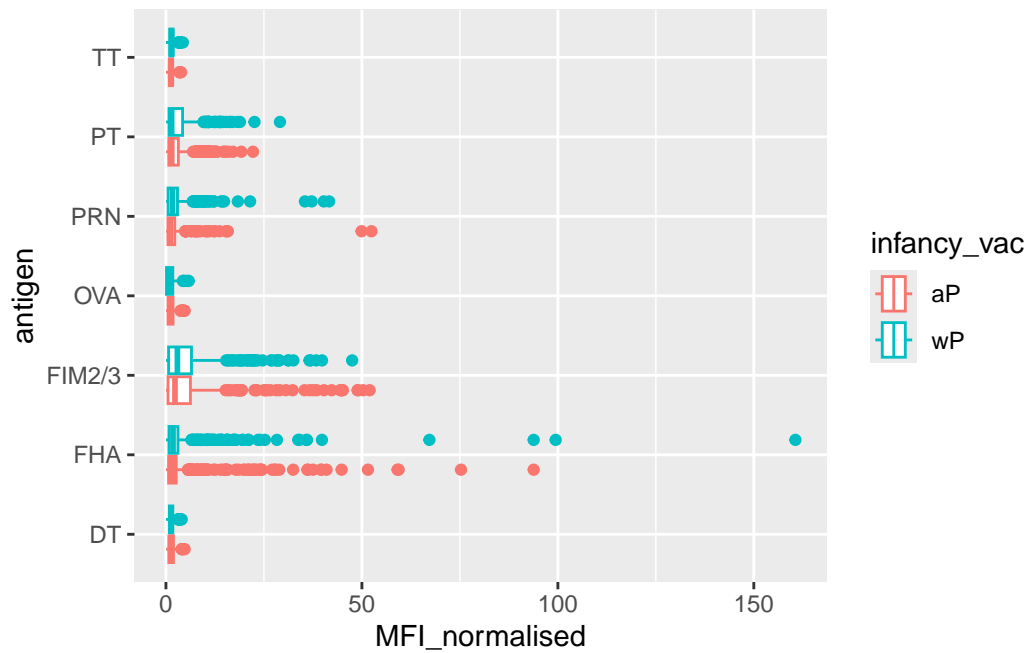
Make a first plot of MFI (mean fluorescence intensity - a measure of how much is detected) for each antigen.

```
ggplot(igg) +
  aes(MFI_normalised, antigen)+
  geom_boxplot()
```



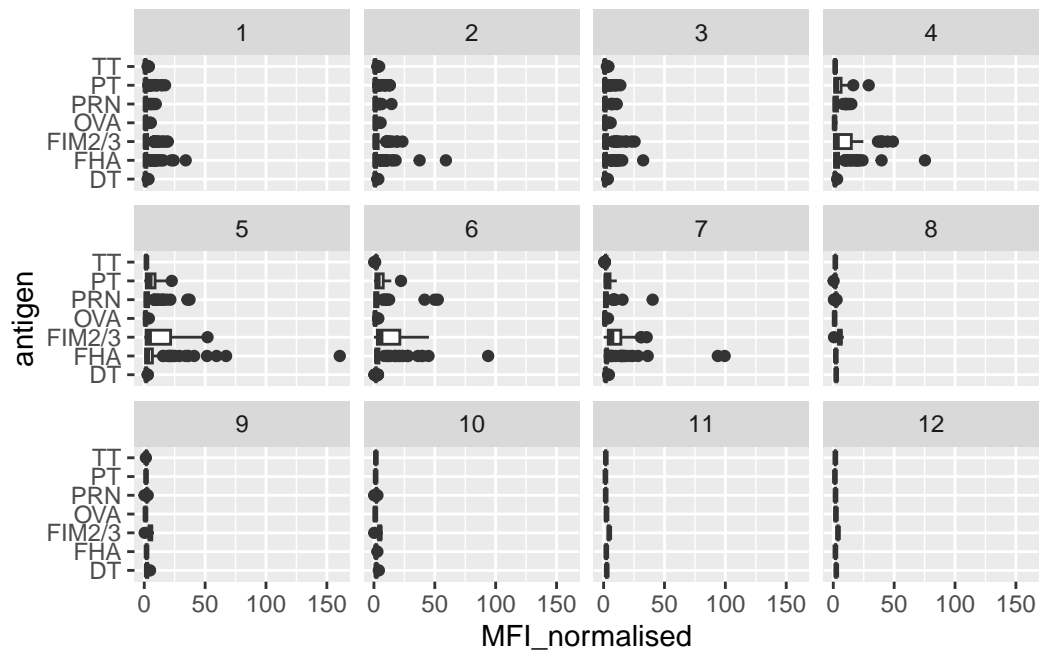
Lets color by aP/wP infancy_vac

```
ggplot(igg) +
  aes(MFI_normalised, antigen, col = infancy_vac)+
  geom_boxplot()
```



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

```
ggplot(igg) +
  aes(MFI_normalised, antigen)+
  geom_boxplot() +
  facet_wrap(~visit)
```



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

PT, PRN, FIM2/3, and FHA shows observable differences in the igG antibody level over time. They are more sensitive to the bacteria.

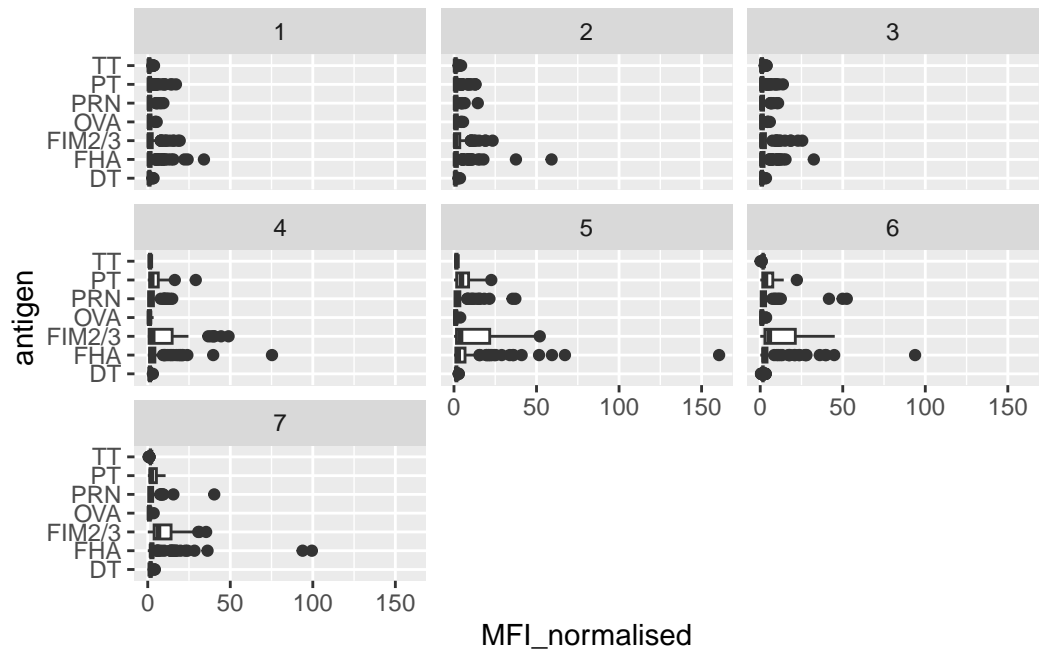
```
table(igg$visit)
```

```
 1  2  3  4  5  6  7  8  9 10 11 12
902 902 930 559 559 540 525 150 147 133 21 21
```

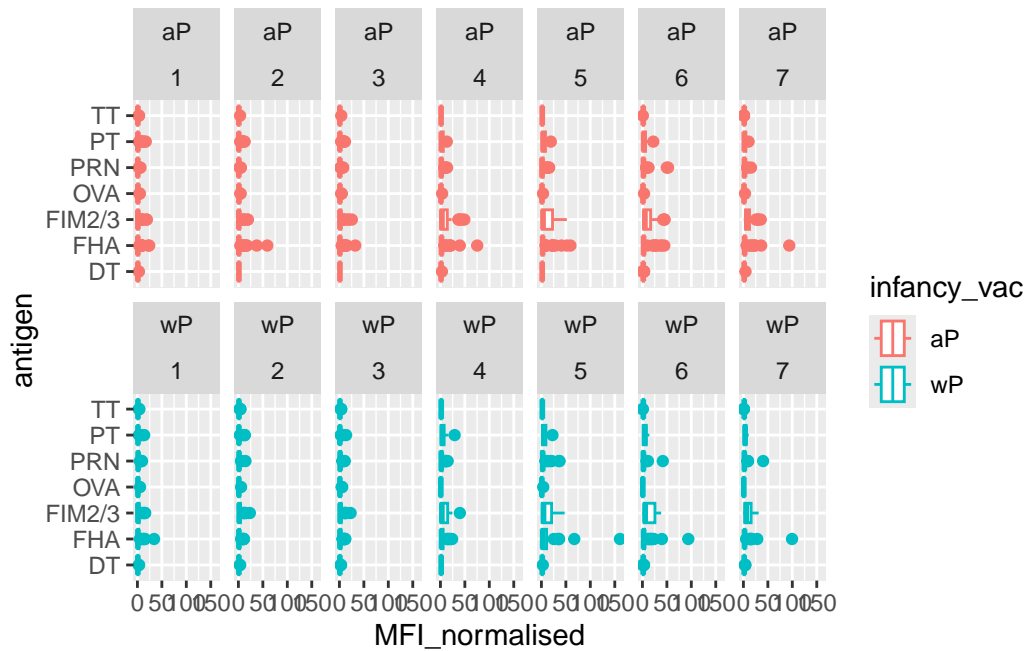
```
igg_7 <- filter(igg, visit %in% 1:7)
table(igg_7$visit)
```

	1	2	3	4	5	6	7
	902	902	930	559	559	540	525

```
ggplot(igg_7) +
  aes(MFI_normalised, antigen)+
  geom_boxplot() +
  facet_wrap(~visit)
```

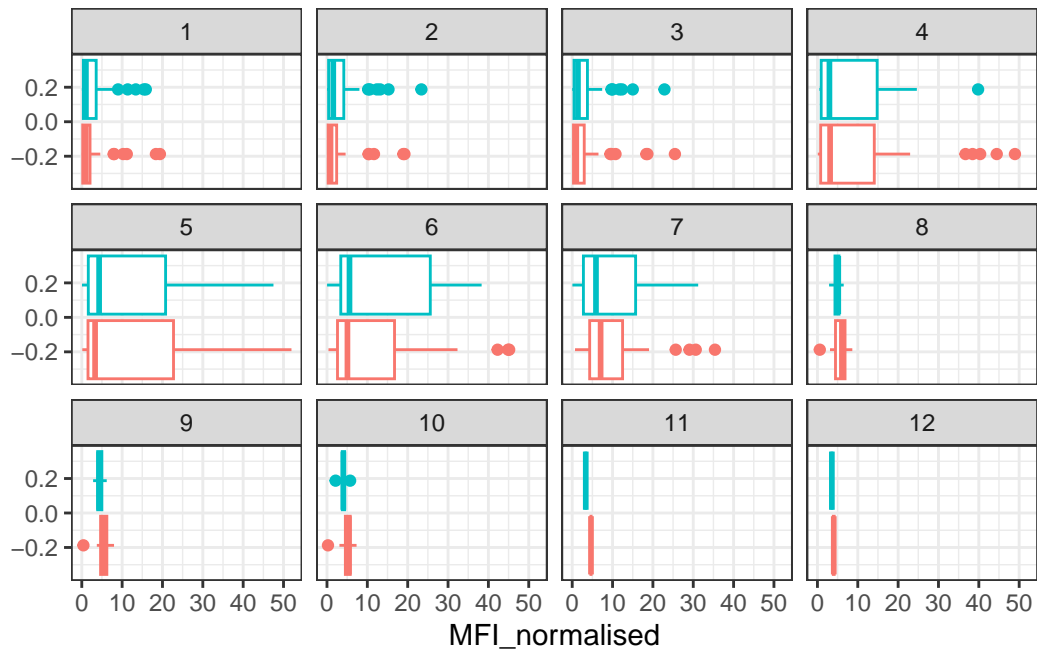


```
ggplot(igg_7) +
  aes(MFI_normalised, antigen, col=infancy_vac)+
  geom_boxplot() +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```

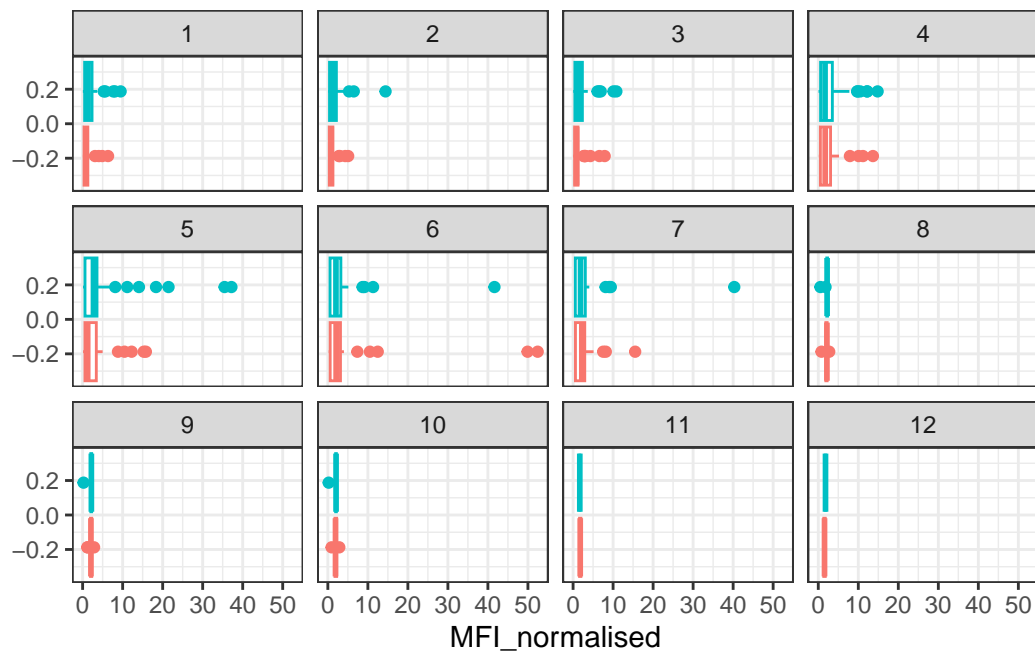



Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a “control” antigen (“OVA”, that is not in our vaccines) and a clear antigen of interest (“PT”, Pertussis Toxin, one of the key virulence factors produced by the bacterium *B. pertussis*).

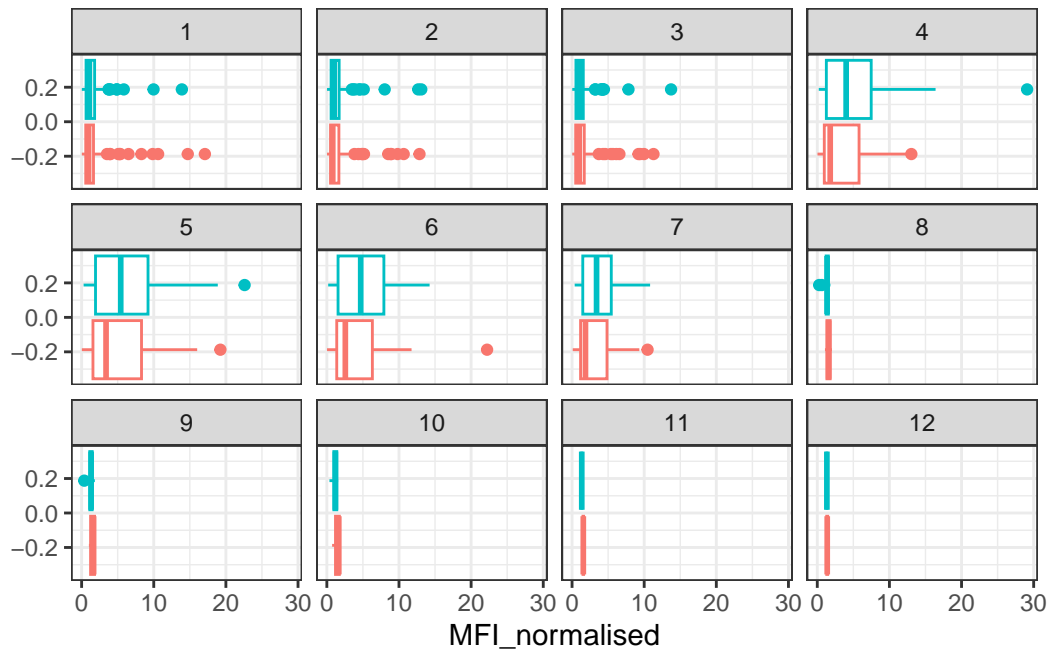
```
filter(igg, antigen=="FIM2/3") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = F) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



```
filter(igg, antigen=="PRN") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = F) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



```
filter(igg, antigen=="PT") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = F) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



Q16. What do you notice about these two antigens time courses and the PT data in particular?

The antigen level clearly rose over time, which is similar with the trend for PT antigen. The antigen level for all three antigens tend to peak at visit 5.

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

This trend is similar for aP and wP subjects, while wP subjects tend to have a larger increase in their antigen level.

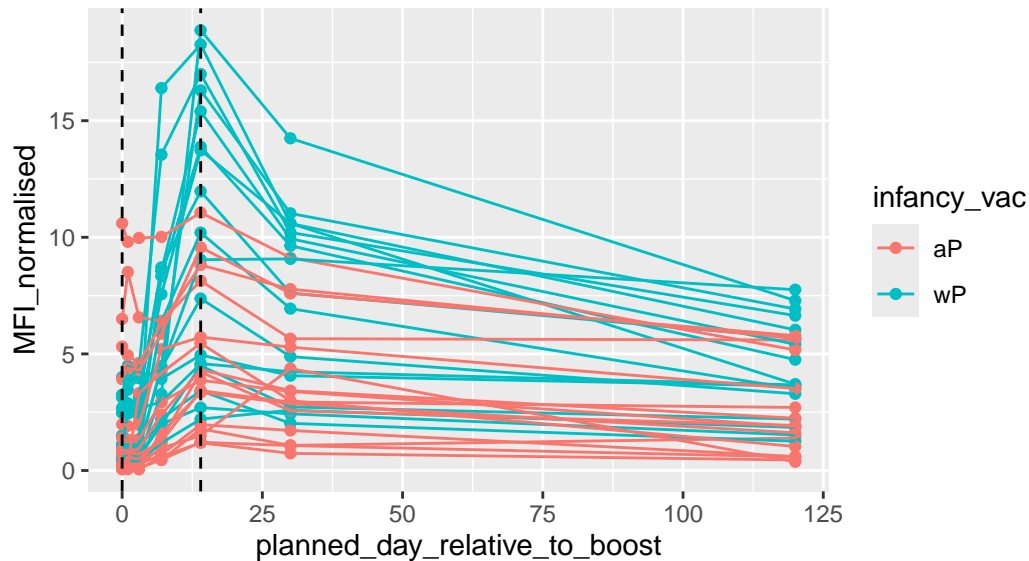
```
abdata.21 <- abdata %>% filter(dataset == "2021_dataset")

abdata.21 %>%
  filter(isotype == "IgG", antigen == "PT") %>%
  ggplot() +
    aes(x=planned_day_relative_to_boost,
         y=MFI_normalised,
         col=infancy_vac,
         group=subject_id) +
    geom_point() +
    geom_line() +
    geom_vline(xintercept=0, linetype="dashed") +
    geom_vline(xintercept=14, linetype="dashed") +
```

```
labs(title="2021 dataset IgG PT",
      subtitle = "Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)")
```

2021 dataset IgG PT

Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)



Q18. Does this trend look similar for the 2020 dataset?

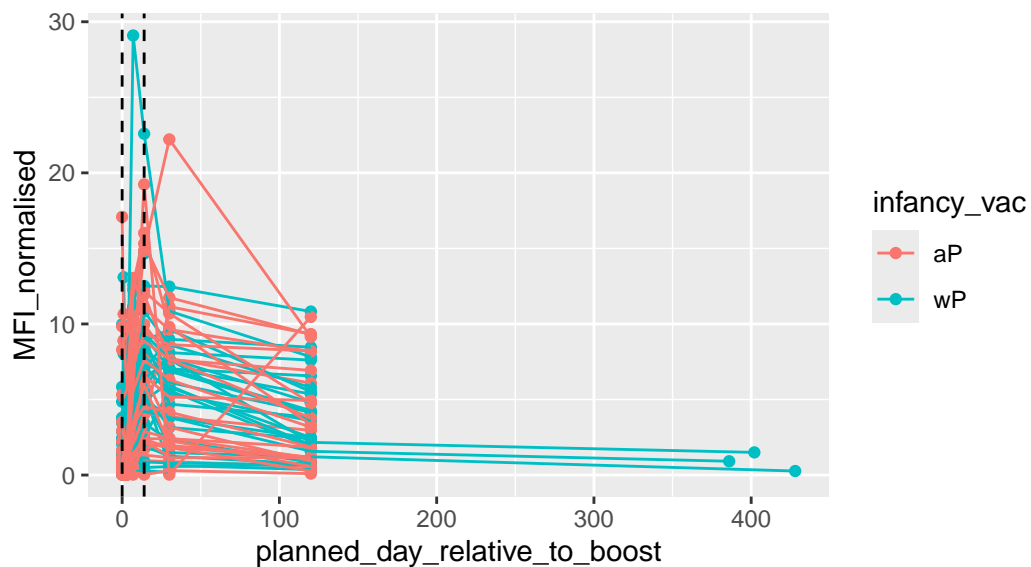
The trend looks different for the 2020 dataset. In the 2020 dataset, the level of antigen for wP tend to decrease comparing to the 2021 dataset.

```
abdata.21 <- abdata %>% filter(dataset == "2020_dataset")

abdata.21 %>%
  filter(isotype == "IgG", antigen == "PT") %>%
  ggplot() +
    aes(x=planned_day_relative_to_boost,
         y=MFI_normalised,
         col=infancy_vac,
         group=subject_id) +
    geom_point() +
    geom_line() +
    geom_vline(xintercept=0, linetype="dashed") +
    geom_vline(xintercept=14, linetype="dashed") +
    labs(title="2020 dataset IgG PT",
          subtitle = "Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)")
```

2020 dataset IgG PT

Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)



Obtaining CMI-PB RNASeq data

IGHG1 gene is the key gene involved in expressing any IgG1 antibody. Let's read available RNA-Seq data for this gene into R and investigate the time course of its gene expression values.

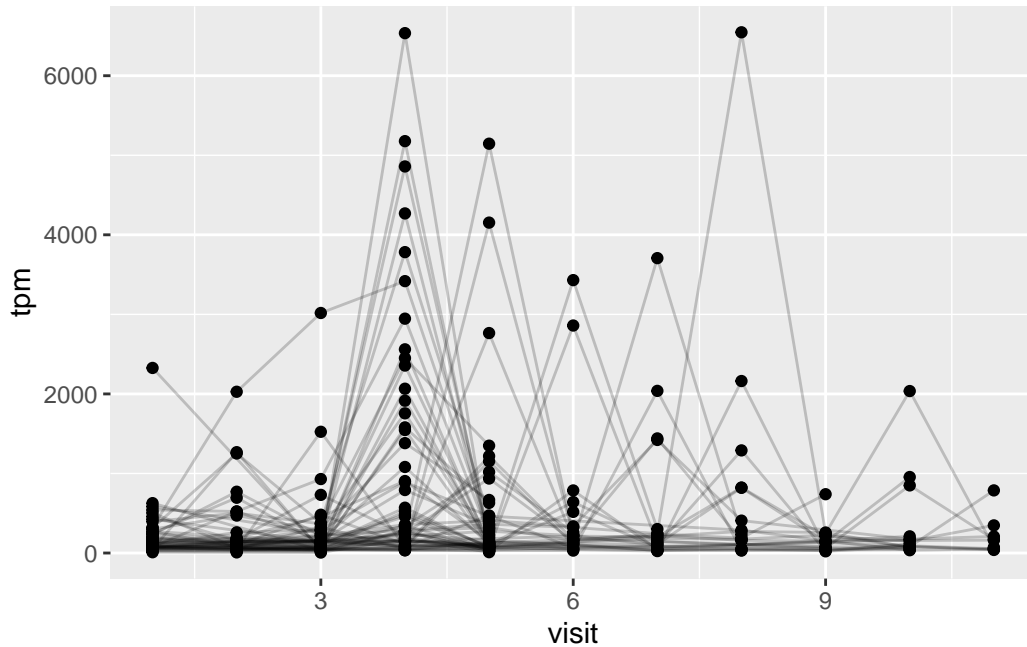
```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSOG00000211896.7"
rna <- read_json(url, simplifyVector = TRUE)
```

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



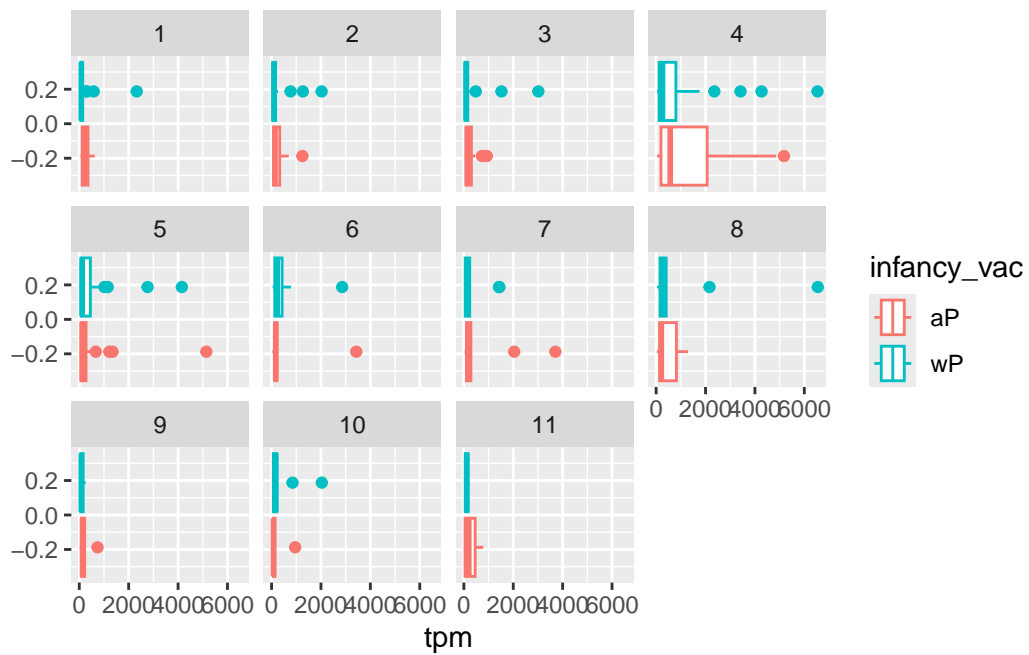
Q20.: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

The expression is at highest during the 4th visit.

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

The antigen level peaks at the 5th visit. The pattern does not match, but it make sense since mRNA transcripts do not instantaneously folds into antigen.

```
ggplot(ssrna) +
  aes(tpm, col=infancy_vac) +
  geom_boxplot() +
  facet_wrap(vars(visit))
```



```
ssrna %>%
  filter(visit==4) %>%
  ggplot() +
    aes(tpm, col=infancy_vac) + geom_density() +
    geom_rug()
```

