

class18

Amit Subramanian

1. Investigating pertussis cases by year

The United States Centers for Disease Control and Prevention (CDC) has been compiling reported pertussis case numbers since 1922 in their National Notifiable Diseases Surveillance System (NNDSS).

We can view this data on the CDC website here: <https://www.cdc.gov/pertussis/surv-reporting/cases-by-year.html>

```
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)
```

```
)
```

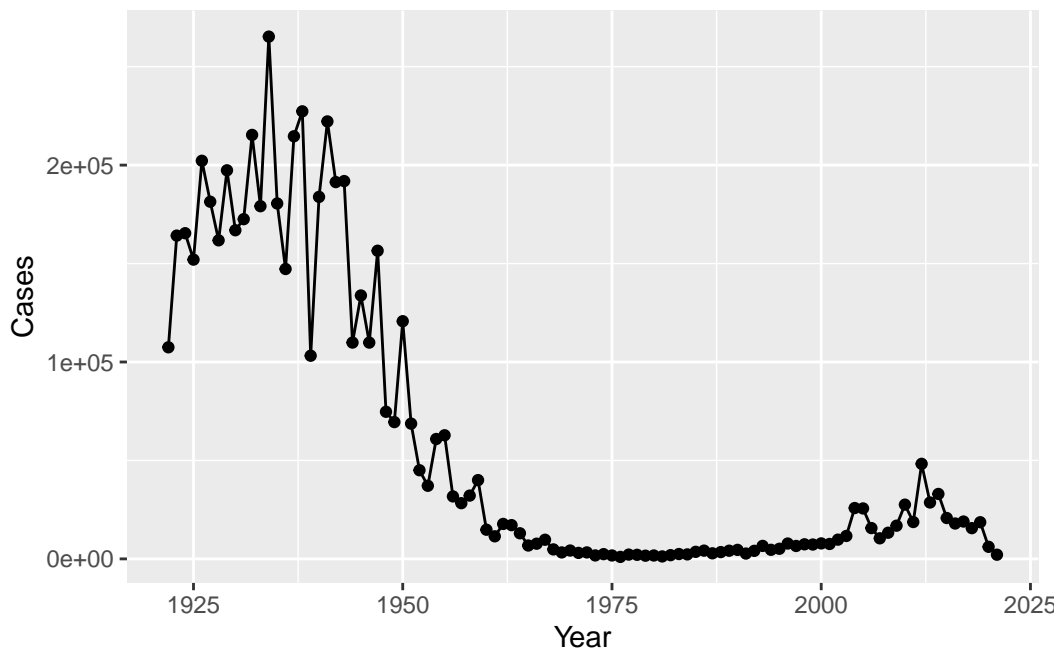
Let's make a plot of case numbers per year.

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)

base <- ggplot(cdc) +
  aes(x = Year, y = Cases) +
  geom_point() +
  geom_line()

base
```



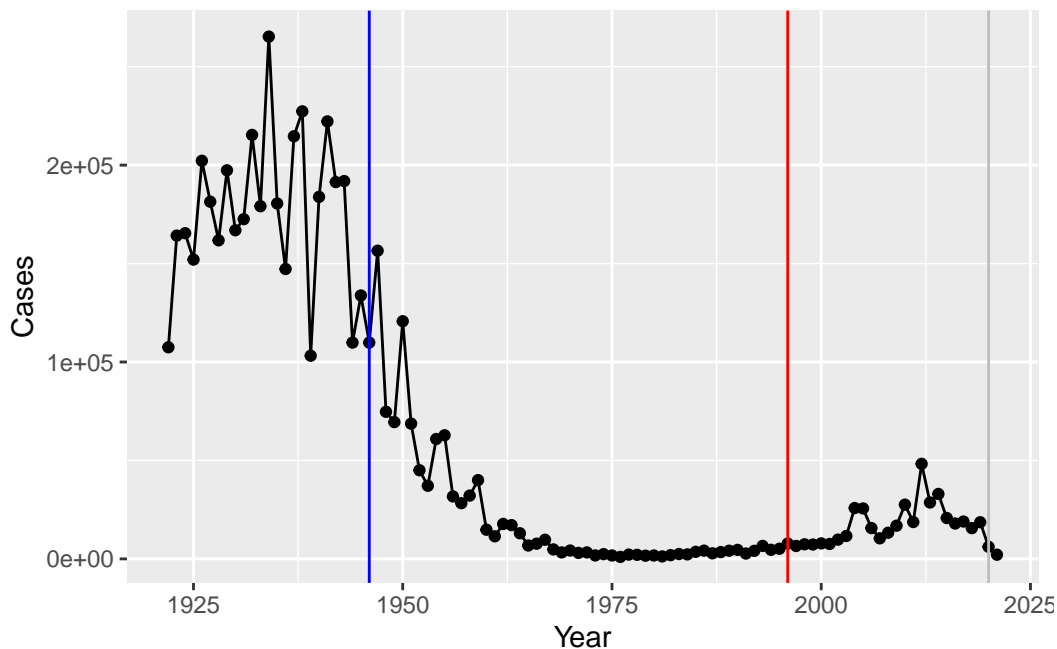
2. A tale of two vaccines (wP & aP)

Two types of pertussis vaccines have been developed: whole-cell pertussis (wP) and acellular pertussis (aP). The first vaccines were composed of ‘whole cell’ (wP) inactivated bacteria. The latter aP vaccines use purified antigens of the bacteria (the most important pertussis components for our immune system).

These aP vaccines were developed to have less side effects than the older wP vaccines and are now the only form administered in the United States.

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?

```
ggplot(cdc) +
  aes(x = Year, y = Cases) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept=1946,col="blue") +
  geom_vline(xintercept=1996, col="red") +
  geom_vline(xintercept=2020, col="grey")
```



I notice that there is a clear drop off and stagnation of a low amount of cases from around 1960-2005.

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.

3. Exploring CMI-PB data

Why is this vaccine-preventable disease on the upswing?

We will use the jsonlite package to read this data into R.

```
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

Let's try to find how many individuals/subjects are in this dataset.

```
nrow(subject)
```

```
[1] 118
```

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

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

```
aP wP
60 58
```

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

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

Female	Male
79	39

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	21	11
Black or African American	2	0
More Than One Race	9	2
Native Hawaiian or Other Pacific Islander	1	1
Unknown or Not Reported	11	4
White	35	20

Side-Note: Working with dates...

```
library(lubridate)
```

Attaching package: 'lubridate'

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

date, intersect, setdiff, union

What is today's date (at the time I am writing this obviously)

```
today()
```

```
[1] "2024-06-11"
```

How many days have passed since new year 2000?

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

Time difference of 8928 days

What is this in years?

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

```
[1] 24.44353
```

Note that here we are using the `ymd()` function to tell lubridate the format of our particular date and then the `time_length()` function to convert days to years.

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?

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

Time differences in days

```
[1] 14041 20616 15137 13311 12215 13311 15867 14406 10389 15502 14041 15502
[13] 10023 11484 12945 13676 16233 10023 11119 15867 15137 14406 12215 11850
[25] 13311 15137 10023 15502 10023 13311 12945 10023 12580 15137 12215 10023
[37] 9658 10023 14406 11119 14406 10023 9658 9658 10023 9658 10389 9658
[49] 10023 10023 10023 9658 9658 10023 10023 10023 10389 10023 10023 10023
[61] 13676 11484 10754 11484 12580 17694 19155 19155 12580 9658 9658 12215
[73] 10754 10754 9658 9658 13311 11484 13676 11850 11484 9658 9293 10023
[85] 8928 9658 8928 8928 10023 9293 9658 8928 10389 9293 9658 8928
[97] 14041 11484 9293 8562 7832 7832 11119 12945 11119 10389 9658 10754
[109] 12945 10023 10389 10389 10389 12580 8197 8928 11119 9658
```

```
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.
21	26	26	27	27	30

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

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
28	32	36	37	39	56

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

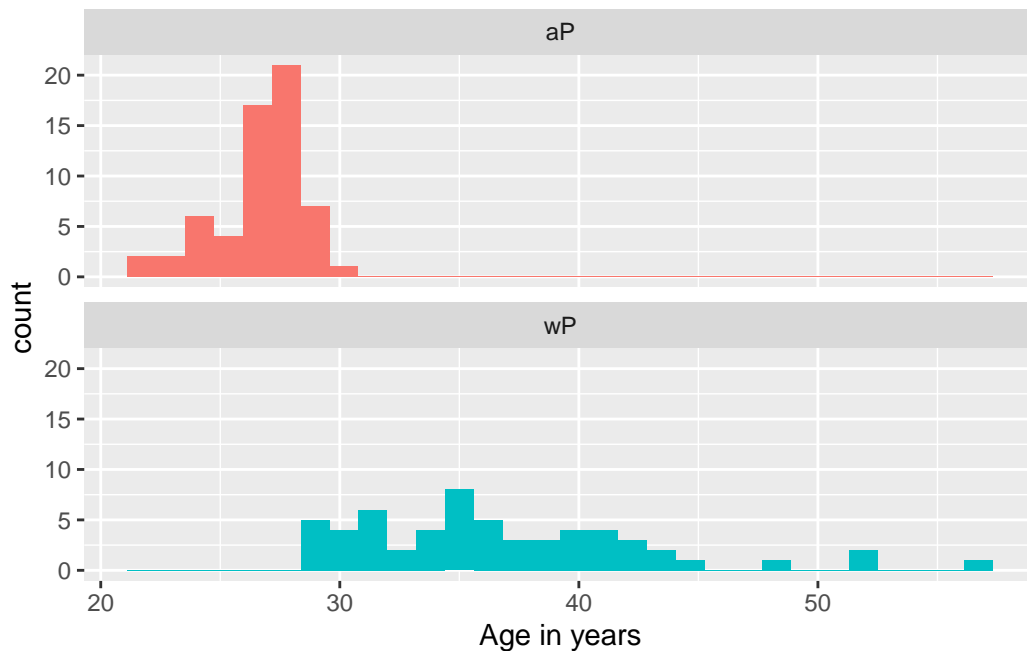
```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)  
age_at_boost <- time_length(int, "year")  
head(age_at_boost)
```

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

Q9. With the help of a faceted boxplot or histogram (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) +  
  xlab("Age in years")
```

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



```
# Or use wilcox.test()
x <- t.test(time_length( wp$age, "years" ),
            time_length( ap$age, "years" ))

x$p.value
```

```
[1] 6.813505e-19
```

Joining multiple tables

```
specimen <- read_json("http://cmi-pb.org/api/specimen", simplifyVector = T)
ab_titer <- read_json("http://cmi-pb.org/api/v4/plasma_ab_titer",simplifyVector = T )
```

```
library(dplyr)
```

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:

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

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

```
dim(meta)
```

```
[1] 939 14
```

```
head(meta)
```

	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	infancy_vac	biological_sex
1	0	Blood	1	wP	Female
2	1	Blood	2	wP	Female
3	3	Blood	3	wP	Female
4	7	Blood	4	wP	Female
5	14	Blood	5	wP	Female
6	30	Blood	6	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
1	14041 days
2	14041 days
3	14041 days
4	14041 days
5	14041 days
6	14041 days

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

Joining with `by = join_by(specimen_id)`

```
dim(abdata)
```

```
[1] 41775    21
```

10. 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.

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

Joining with `by = join_by(specimen_id)`

```
dim(abdata)
```

```
[1] 41775    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 3233 7961 7961 7961 7961
```

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

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

```
2020_dataset 2021_dataset 2022_dataset
        31520         8085         2170
```

4. Examine IgG Ab titer levels

Now let's use our joined/merged/linked abdata dataset filter() for IgG isotype.

```
igg <- abdata %>% filter(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	actual_day_relative_to_boost
1	IU/ML	0.530000	1	-3
2	IU/ML	6.205949	1	-3
3	IU/ML	4.679535	1	-3
4	IU/ML	0.530000	3	-3
5	IU/ML	6.205949	3	-3
6	IU/ML	4.679535	3	-3

	planned_day_relative_to_boost	specimen_type	visit	infancy_vac	biological_sex
1	0	Blood	1	wP	Female
2	0	Blood	1	wP	Female
3	0	Blood	1	wP	Female
4	0	Blood	1	wP	Female
5	0	Blood	1	wP	Female
6	0	Blood	1	wP	Female

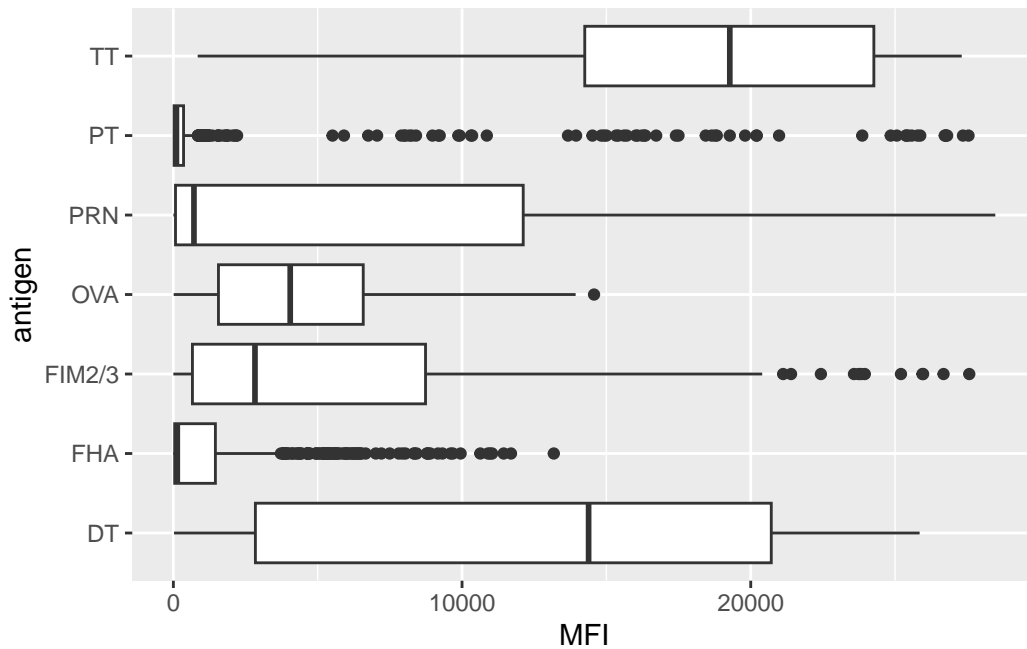
	ethnicity	race	year_of_birth	date_of_boost	dataset
1	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
2	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
3	Not Hispanic or Latino	White	1986-01-01	2016-09-12	2020_dataset
4	Unknown	White	1983-01-01	2016-10-10	2020_dataset
5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

	age
1	14041 days
2	14041 days
3	14041 days
4	15137 days
5	15137 days
6	15137 days

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

```
base <- ggplot(igg) +
  aes(MFI, antigen) +
  geom_boxplot()
```

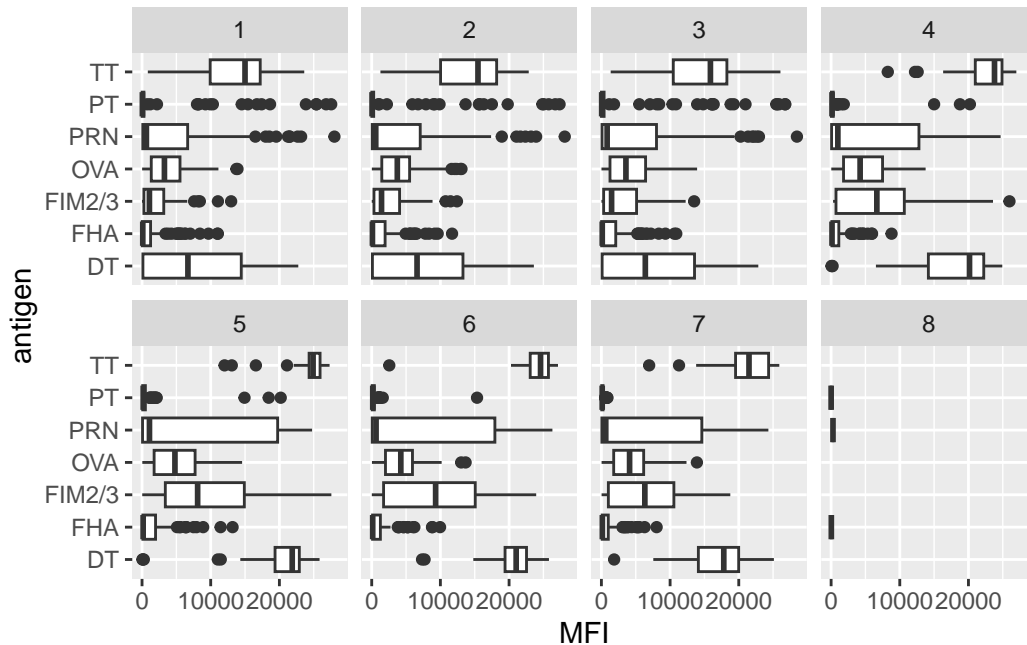
```
base
```



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

```
 1  2  3  4  5  6  7  8
524 531 552 426 426 393 378  3
```

```
base +
  facet_wrap(vars(visit), nrow=2)
```



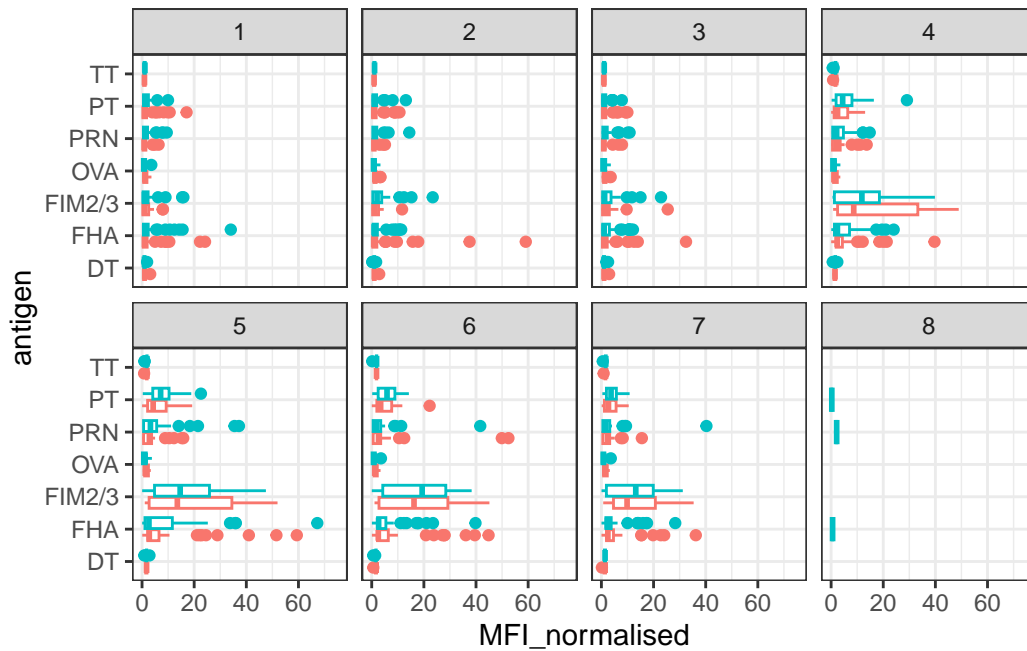
Q14. What antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others?

You can use the CMI-PB website search functionality and Terminology Browser (under development) to find out about each antigen. Note that this is still work in progress.

We can attempt to examine differences between wP and aP here by setting color and/or facet values of the plot to include `infancy_vac` status (see below). However these plots tend to be rather busy and thus hard to interpret easily.

```
ggplot(igg) +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  xlim(0,75) +
  theme_bw()
```

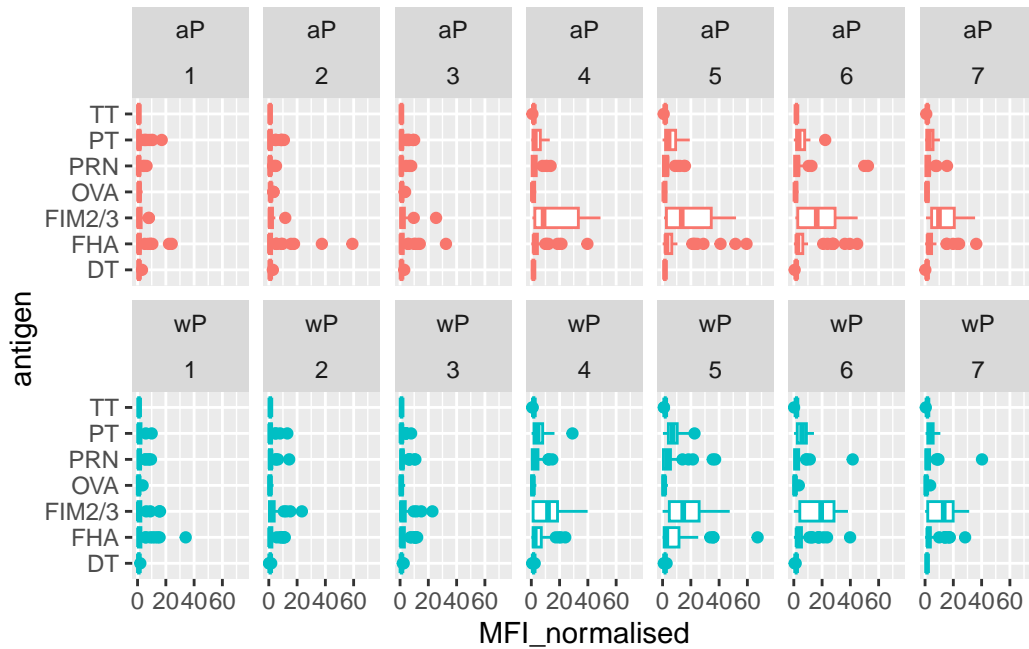
Warning: Removed 5 rows containing non-finite outside the scale range (``stat_boxplot()``).



Another version of this plot adding infancy_vac to the faceting:

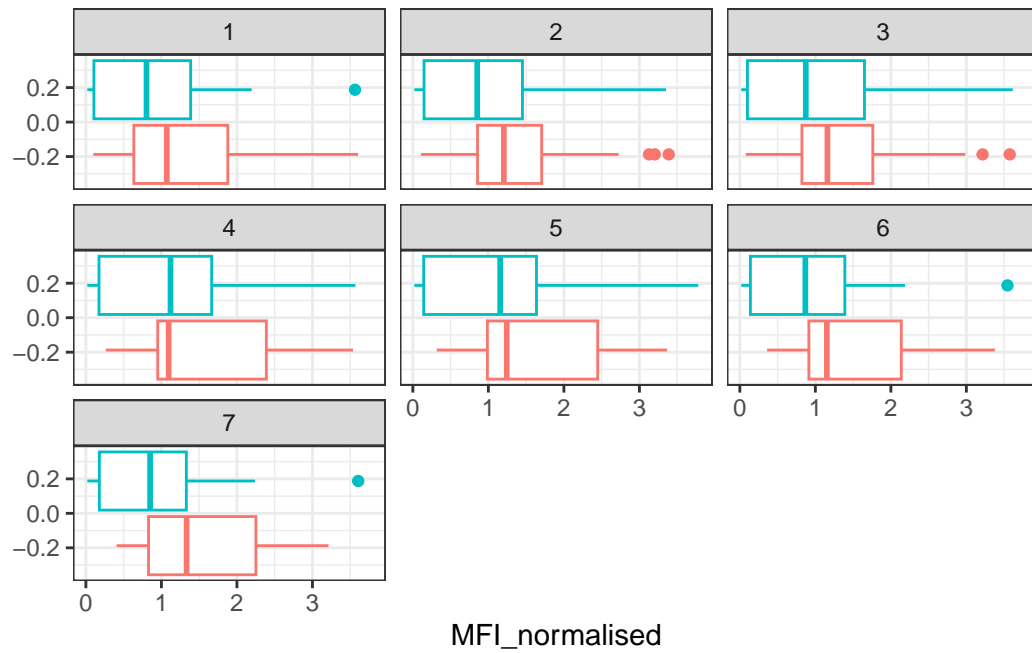
```
igg %>% filter(visit != 8) %>%
ggplot() +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  xlim(0,75) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```

Warning: Removed 5 rows containing non-finite outside the scale range (`stat_boxplot()`).

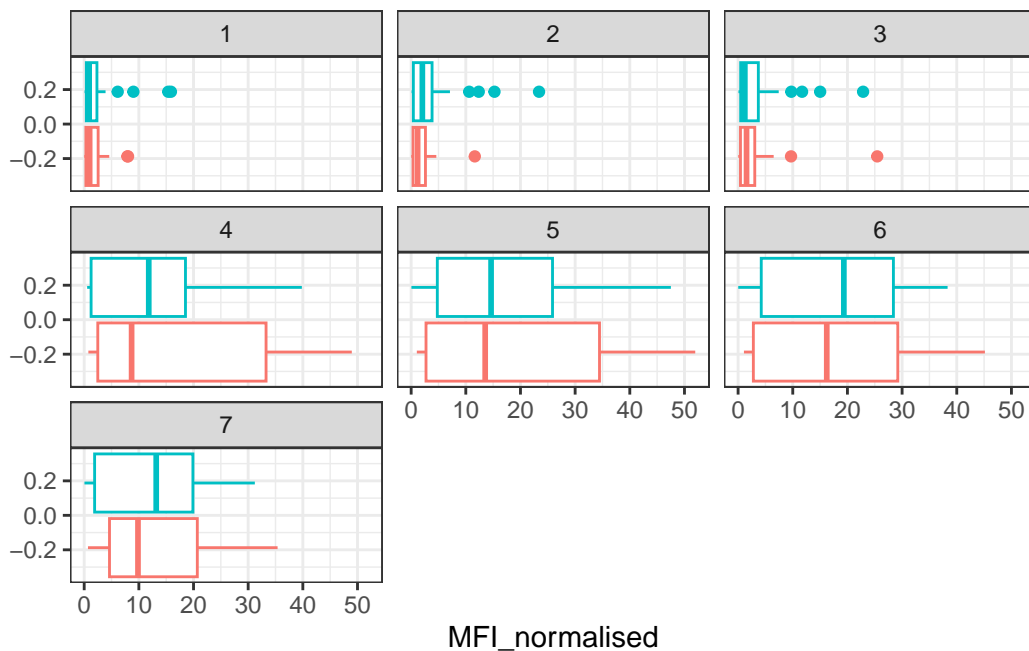


Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a “control” antigen (“OVA”, that is not in our vaccines) and a clear antigen of interest (“PT”, Pertussis Toxin, one of the key virulence factors produced by the bacterium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

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

```
filter(igg, antigen=="FIM2/3") %>%
  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?

PT levels clearly rise over time and far exceed those of OVA. They also appear to peak at visit 5 and then decline. This trend appears similar for wP and aP subjects.

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

I noticed that aP appeared to respond slower than wP did.

Lets finish this section by looking at the 2021 dataset IgG PT antigen levels time-course:

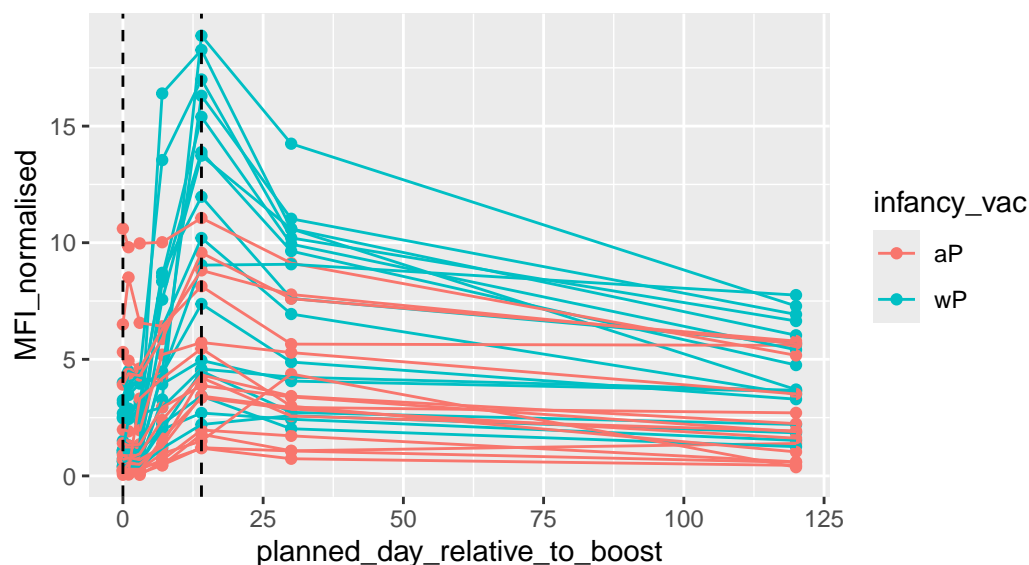
```
abdata.21 <- abdata %>% filter(dataset == "2021_dataset")

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)



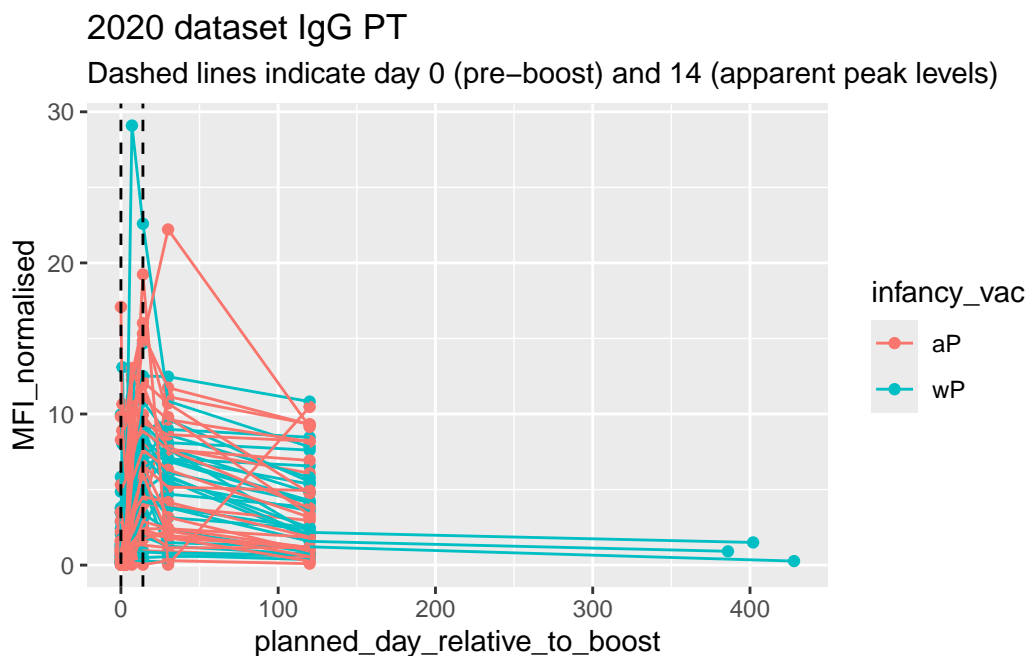
Q18. Does this trend look similar for the 2020 dataset?

Yes, this trend look similar for the 2020 dataset.

```
abdata.20 <- abdata %>% filter(dataset == "2020_dataset")

abdata.20 %>%
  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)")
```



5. Obtaining CMI-PB RNASeq data

For RNA-Seq data the API query mechanism quickly hits the web browser interface limit for file size. We will present alternative download mechanisms for larger CMI-PB datasets in the next section. However, we can still do “targeted” RNA-Seq queries via the web accessible API.

The link above is for the key gene involved in expressing any IgG1 antibody, namely the IGHG1 gene. Let’s read available RNA-Seq data for this gene into R and investigate the time course of it’s gene expression values.

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

Let’s take a look at this genes TPM expression values over aP/wP status and at different visits.

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

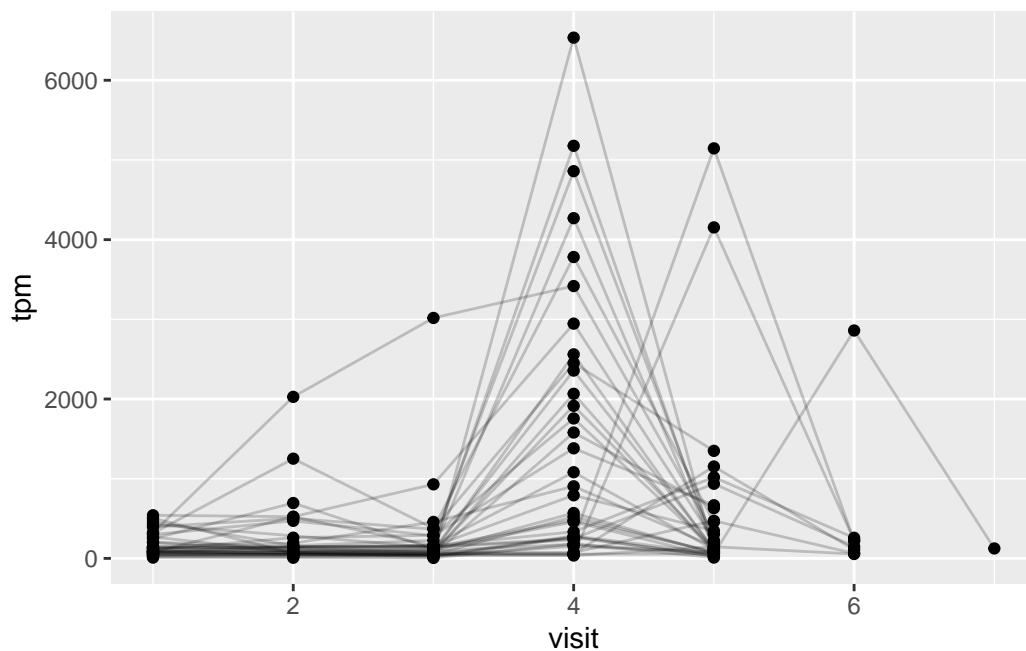
Joining with `by = join_by(subject_id)`

```
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 its maximum level)?

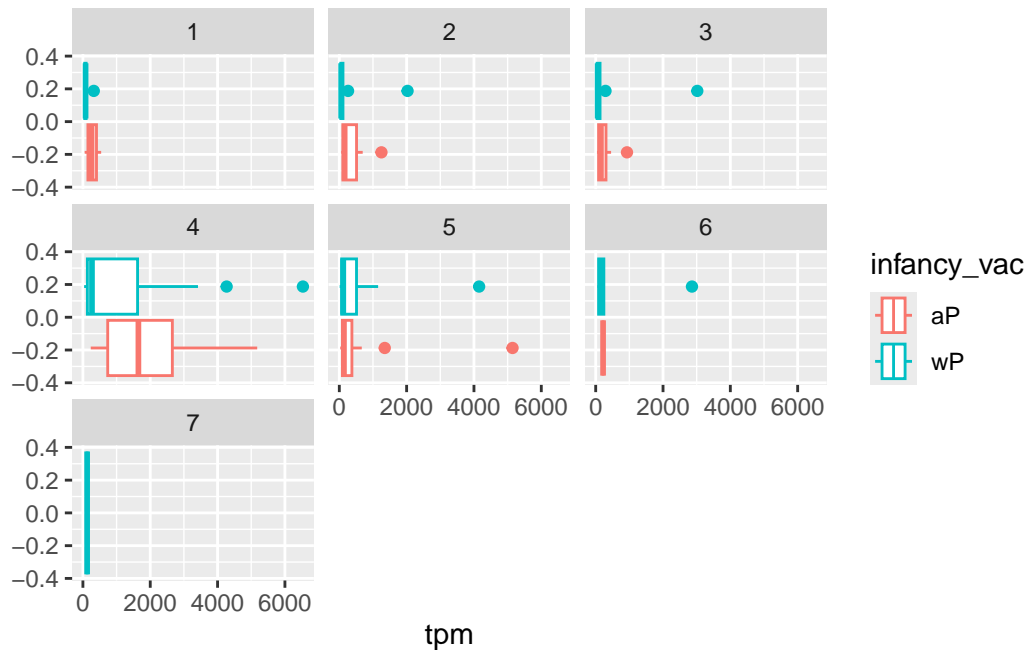
The expression of this gene is at its max level at visit 4.

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

Yes, it does match the trend of antibody titer data.

Let's dig deeper and color and/or facet by infancy_vac status:

```
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()
```

