Pertussis Mini Project

Hetian Su

Q1

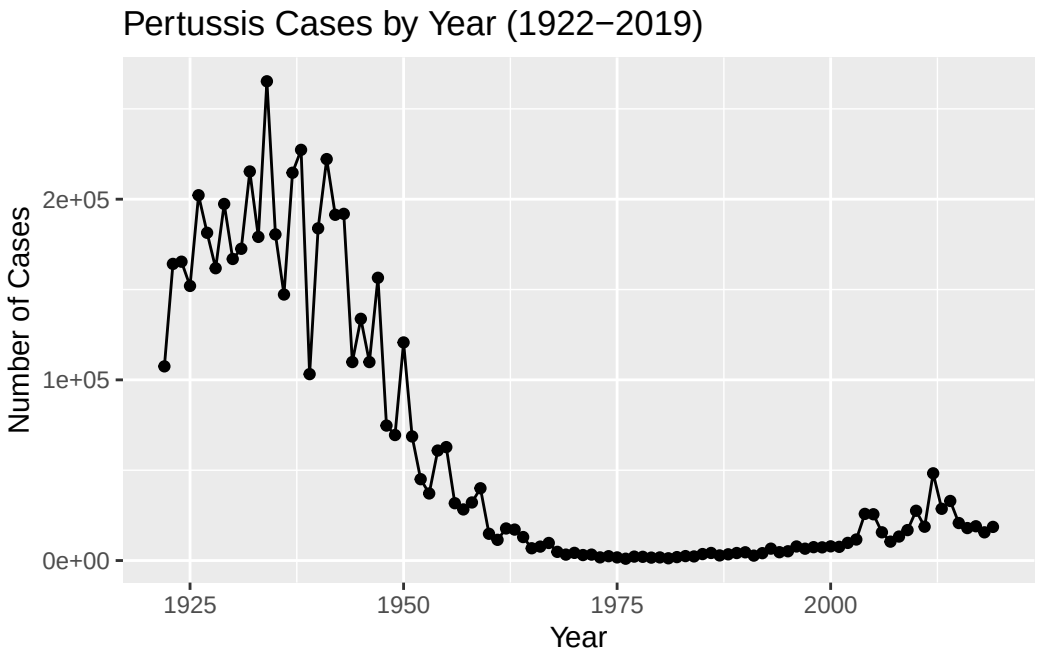
```
# install.packages('datapasta')
library(datapasta)
```

```
Warning: package 'datapasta' was built under R version 4.2.2
```

```
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),
  No..Reported.Pertussis.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,
```

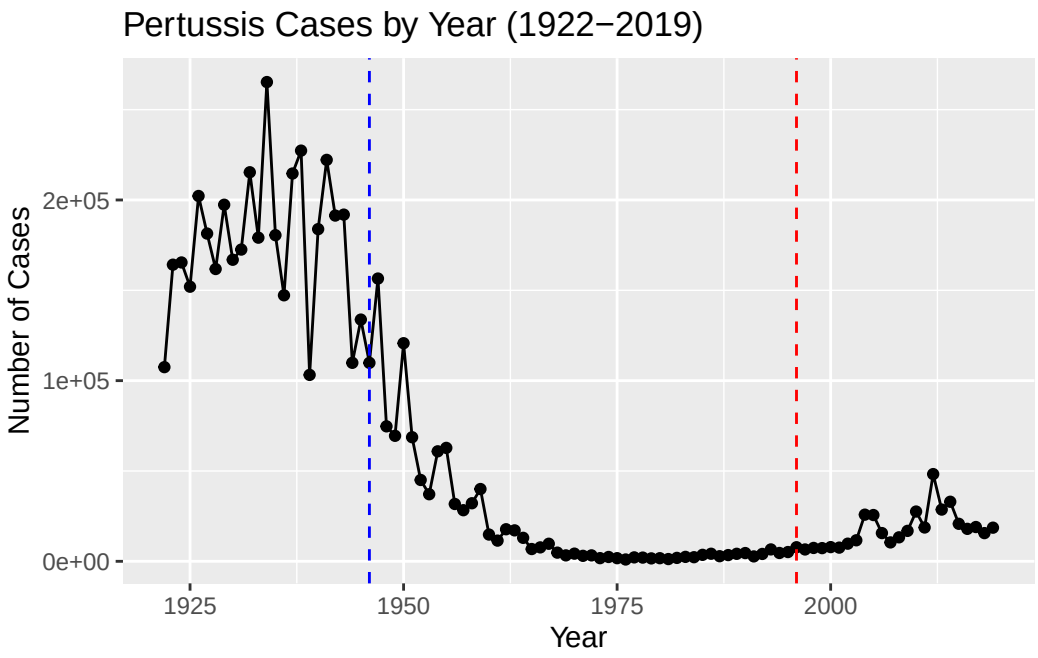
```
                                68687,45030,37129,60886,62786,31732,28295,
                                32148,40005,14809,11468,17749,17135,
                                13005,6799,7717,9718,4810,3285,4249,
                                3036,3287,1759,2402,1738,1010,2177,2063,
                                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)
)
```

```
library(ggplot2)
plot <- ggplot(cdc)+
  aes(Year, No..Reported.Pertussis.Cases)+
  geom_point()+
  geom_line()+
  labs(title='Pertussis Cases by Year (1922-2019)', x='Year', y='Number of Cases')
plot
```



Q2

```
plot <- plot + geom_vline(xintercept = 1946, color='blue', linetype='dashed') + geom_vline
plot
```



Q3

It can be seen that after the introduction of aP vaccine, the number of cases started to increase again. It is possible that the bacterial evolution could more easily escape the protection provided by the aP vaccine.

```
library(jsonlite)
```

```
Warning: package 'jsonlite' was built under R version 4.2.2
```

```
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

```
sum(subject$infancy_vac=='wP')
```

```
[1] 49
```

```
sum(subject$infancy_vac=='aP')
```

```
[1] 47
```

There are 47 aP vaccinated subjects, and 49 wP vaccinated subjects.

Q5

```
sum(subject$biological_sex=='Male')
```

```
[1] 30
```

```
sum(subject$biological_sex=='Female')
```

```
[1] 66
```

There are 30 males and 66 females.

Q6

```
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
```

```
# install.packages('lubridate')
library(lubridate)
```

```
Warning: package 'lubridate' was built under R version 4.2.2
```

```
Loading required package: timechange
```

```
Warning: package 'timechange' was built under R version 4.2.2
```

```

Attaching package: 'lubridate'
```

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

    date, intersect, setdiff, union
```

Q7

```
wP_age <- time_length(today()-ymd(subject$year_of_birth[subject$infancy_vac=='wP']), 'year
aP_age <- time_length(today()-ymd(subject$year_of_birth[subject$infancy_vac=='aP']), 'year
```

```
mean(wP_age)
```

```
[1] 36.07532
```

```r
mean(aP_age)
```

```
[1] 25.23087
```

```r
t.test(wP_age, aP_age)
```

```
    Welch Two Sample t-test

data:  wP_age and aP_age
t = 12.092, df = 51.082, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  9.044045 12.644857
sample estimates:
mean of x mean of y
 36.07532  25.23087
```

The mean age of wP vaccinated subjects is 36 years old, and that of aP vaccinated subjects is 25 years old. They are significantly different under 2-sample t test.

Q8

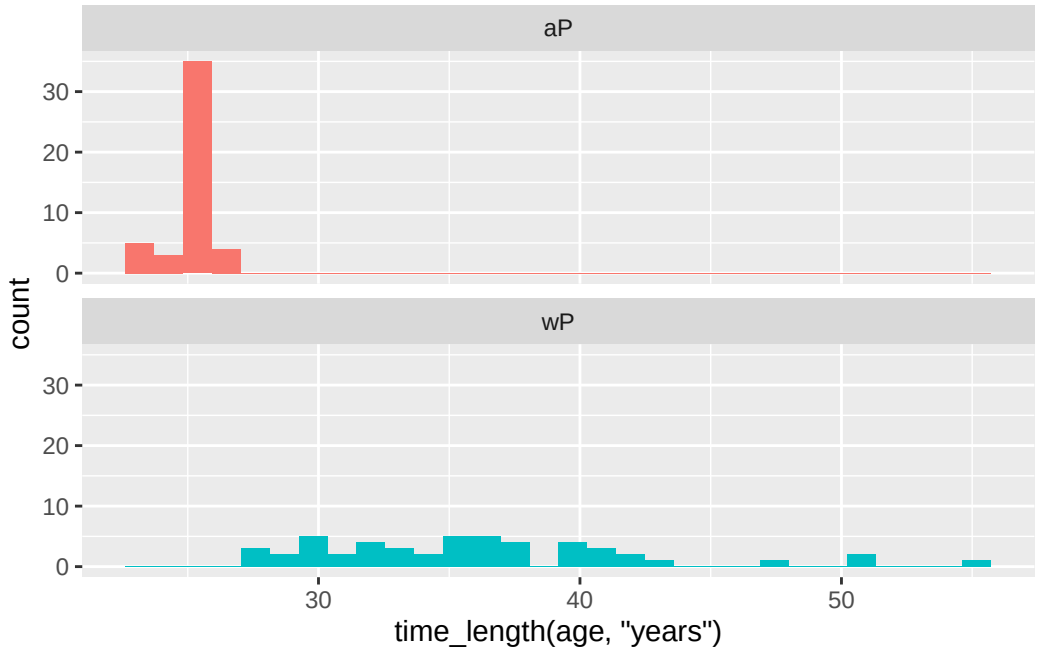
```r
age_at_boost <- time_length(ymd(subject$date_of_boost)-ymd(subject$year_of_birth), 'years'
head(age_at_boost)
```

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

Q9.a

```r
subject$age <- today()-ymd(subject$year_of_birth)
ggplot(subject)+
  aes(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`.
```



The distribution shown in the plot here shows that the 2 groups have significantly different ages.

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

Q9.b

```
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(specimen, subject)
```

```
Joining, by = "subject_id"
```

```
dim(meta)
```

```
[1] 729  14
```

```
head(meta)
```

```
  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 infancy_vac biological_sex
1                             0         Blood     1          wP         Female
2                           736         Blood    10          wP         Female
3                             1         Blood     2          wP         Female
4                             3         Blood     3          wP         Female
5                             7         Blood     4          wP         Female
6                            14         Blood     5          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 13482 days
2 13482 days
```

```
3 13482 days
4 13482 days
5 13482 days
6 13482 days
```

Q10

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

```
Joining, by = "specimen_id"
```

```
dim(abdata)
```

```
[1] 32675    21
```

Q11

```
table(abdata$isotype)
```

```
 IgE  IgG IgG1 IgG2 IgG3 IgG4
6698 1413 6141 6141 6141 6141
```

Q12

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

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

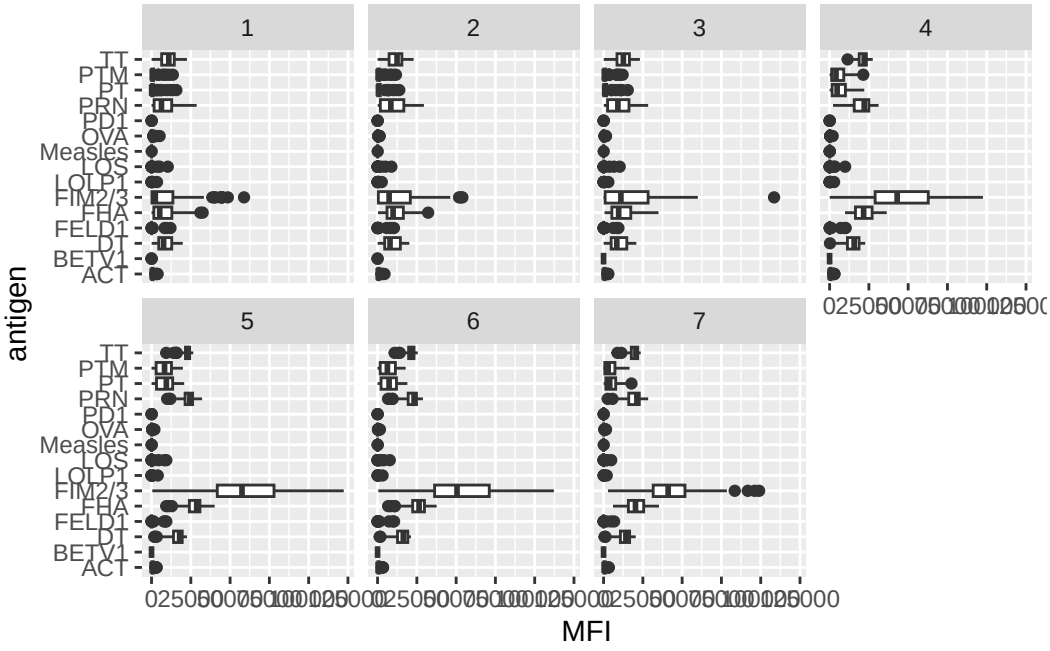
There are a lot fewer specimens collected on the 8th visit.

```
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 actual_day_relative_to_boost
1 IU/ML                3.848750          1                           -3
2 IU/ML                4.357917          1                           -3
3 IU/ML                2.699944          1                           -3
4 IU/ML                1.734784          1                           -3
5 IU/ML                2.550606          1                           -3
6 IU/ML                4.438966          1                           -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 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 13482 days
2 13482 days
3 13482 days
4 13482 days
5 13482 days
6 13482 days
```

Q13

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



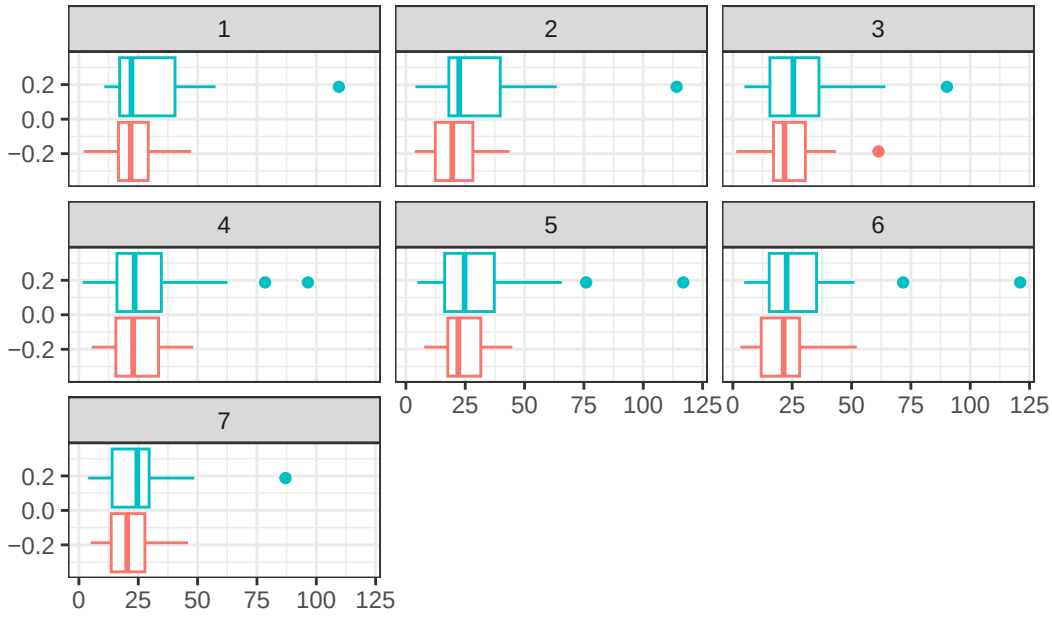
Q14

Judging by the change of mean and distribution shown by boxplots over visits, TT, PRN, FIM2/3, FHA levels increased overtime, with FIM2/3 being the most significant.

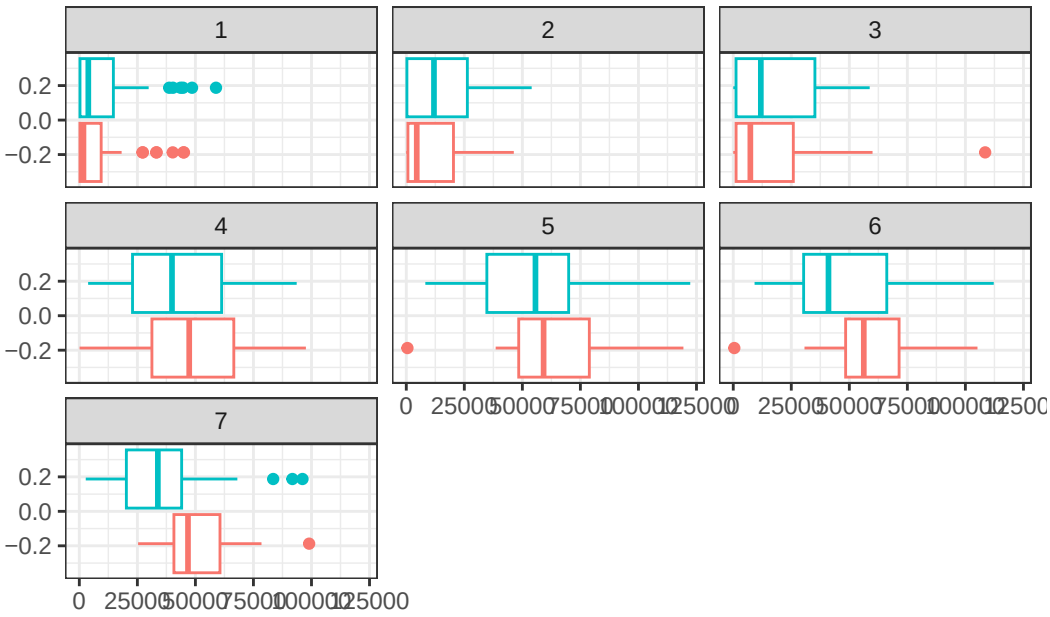
Q15

```
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 = FALSE)+
  facet_wrap(vars(visit))+
  theme_bw()
```



Q16

As compared to the negative control, both wP and aP vaccine induced high level production of FIM2/3. Also notably, overtime the level of FIM2/3 induced by aP maintains but that by wP decreases.

Q17

Overtime, antigen produced in response to aP increases and stays at high levels, thus eventually exceeding the level of that induced by wP which decreases after some time.

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

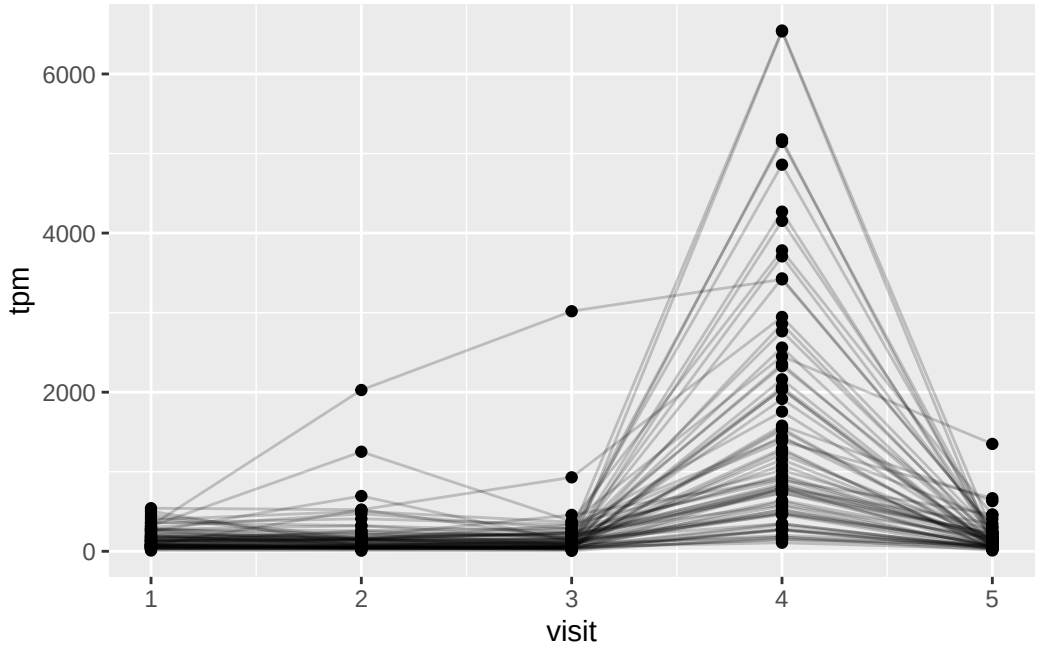
```
rna <- read_json(url, simplifyVector = TRUE)
```

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

```
Joining, by = "specimen_id"
```

Q18

```
ggplot(ssrna)+
  aes(visit, tpm, group=subject_id)+
  geom_point()+
  geom_line(alpha=0.2)
```



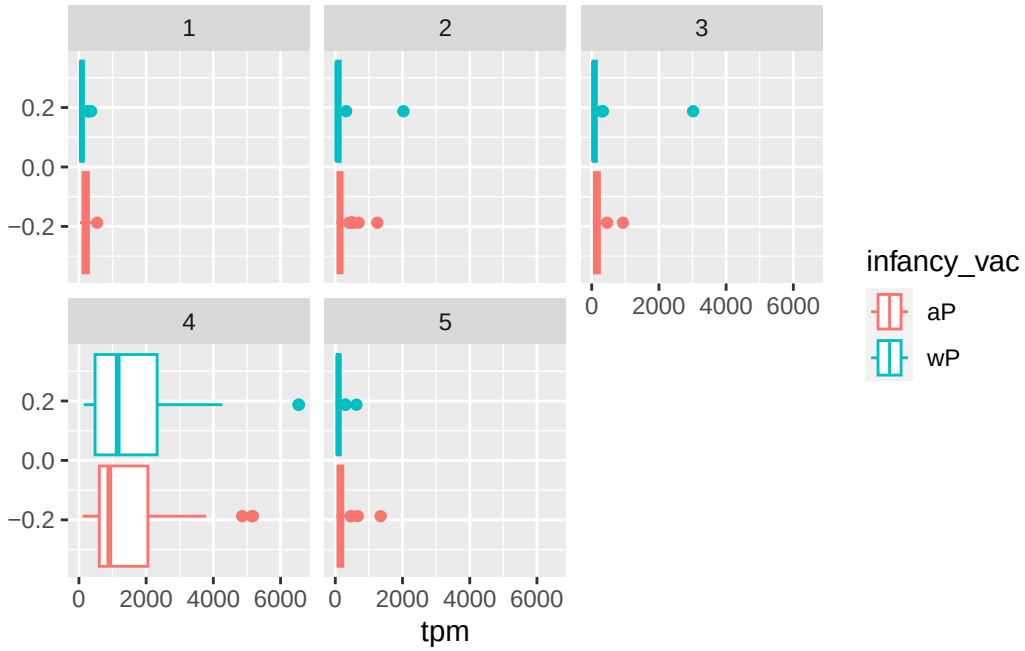
Q19

The expression of this gene tends to increase abruptly and peak at the 4th visit.

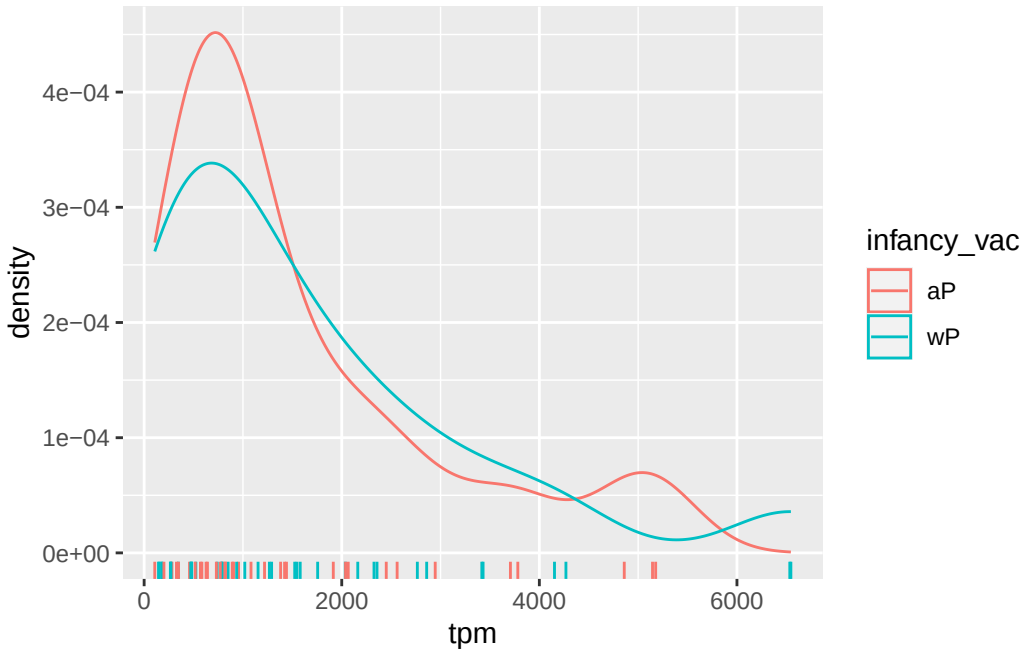
Q20

The peaking behavior matches closely with FIM2/3 pattern induced by wP, whereas FIM2/3 induced by aP seem to be able to stay at rather high levels.

```
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()
```


density
4e−04
3e−04
2e−04
1e−04
0e+00
0
2000
4000
6000
tpm
infancy_vac
aP
wP