Class 18: Pertussis and the CMI-PB Project

Arshiya Zarmahd (PID: A16247996)

1. Investigating Pertussis Cases By 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.

```
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),
  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,
                                    32971,20762,17972,18975,15609,18617,6124,
                                    2116)
)
cdc
```

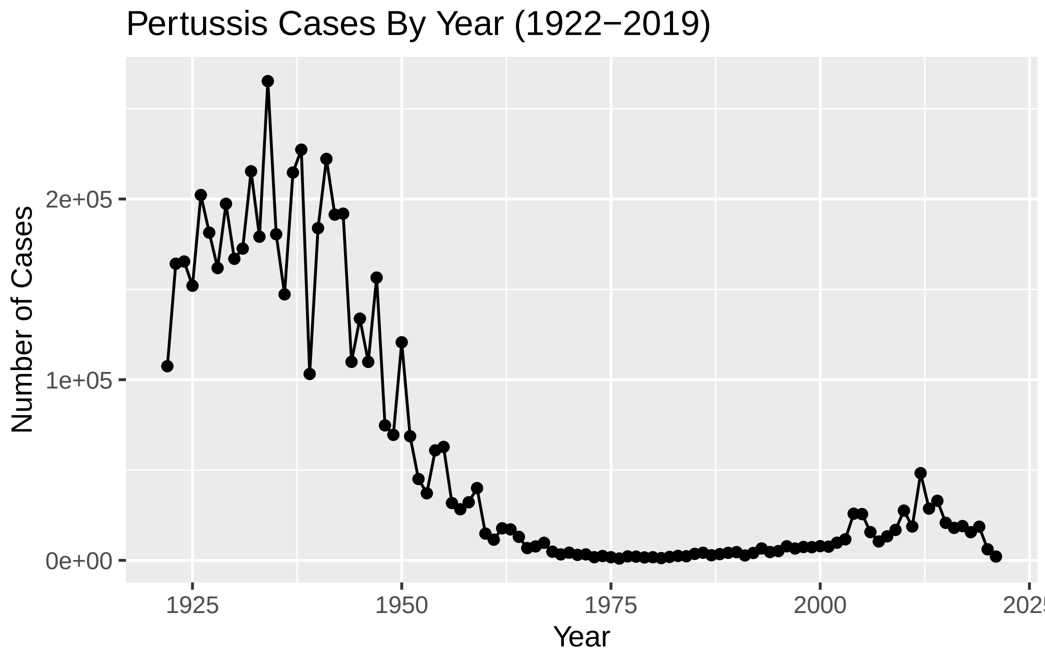
```
   Year No..Reported.Pertussis.Cases
1  1922                       107473
2  1923                       164191
3  1924                       165418
4  1925                       152003
5  1926                       202210
6  1927                       181411
7  1928                       161799
8  1929                       197371
9  1930                       166914
10 1931                       172559
11 1932                       215343
12 1933                       179135
13 1934                       265269
14 1935                       180518
15 1936                       147237
16 1937                       214652
17 1938                       227319
18 1939                       103188
19 1940                       183866
20 1941                       222202
21 1942                       191383
22 1943                       191890
23 1944                       109873
24 1945                       133792
25 1946                       109860
26 1947                       156517
27 1948                        74715
28 1949                        69479
```

29	1950	120718
30	1951	68687
31	1952	45030
32	1953	37129
33	1954	60886
34	1955	62786
35	1956	31732
36	1957	28295
37	1958	32148
38	1959	40005
39	1960	14809
40	1961	11468
41	1962	17749
42	1963	17135
43	1964	13005
44	1965	6799
45	1966	7717
46	1967	9718
47	1968	4810
48	1969	3285
49	1970	4249
50	1971	3036
51	1972	3287
52	1973	1759
53	1974	2402
54	1975	1738
55	1976	1010
56	1977	2177
57	1978	2063
58	1979	1623
59	1980	1730
60	1981	1248
61	1982	1895
62	1983	2463
63	1984	2276
64	1985	3589
65	1986	4195
66	1987	2823
67	1988	3450
68	1989	4157
69	1990	4570
70	1991	2719
71	1992	4083

```
72  1993                        6586
73  1994                        4617
74  1995                        5137
75  1996                        7796
76  1997                        6564
77  1998                        7405
78  1999                        7298
79  2000                        7867
80  2001                        7580
81  2002                        9771
82  2003                       11647
83  2004                       25827
84  2005                       25616
85  2006                       15632
86  2007                       10454
87  2008                       13278
88  2009                       16858
89  2010                       27550
90  2011                       18719
91  2012                       48277
92  2013                       28639
93  2014                       32971
94  2015                       20762
95  2016                       17972
96  2017                       18975
97  2018                       15609
98  2019                       18617
99  2020                        6124
100 2021                        2116
```

```r
library(ggplot2)
```

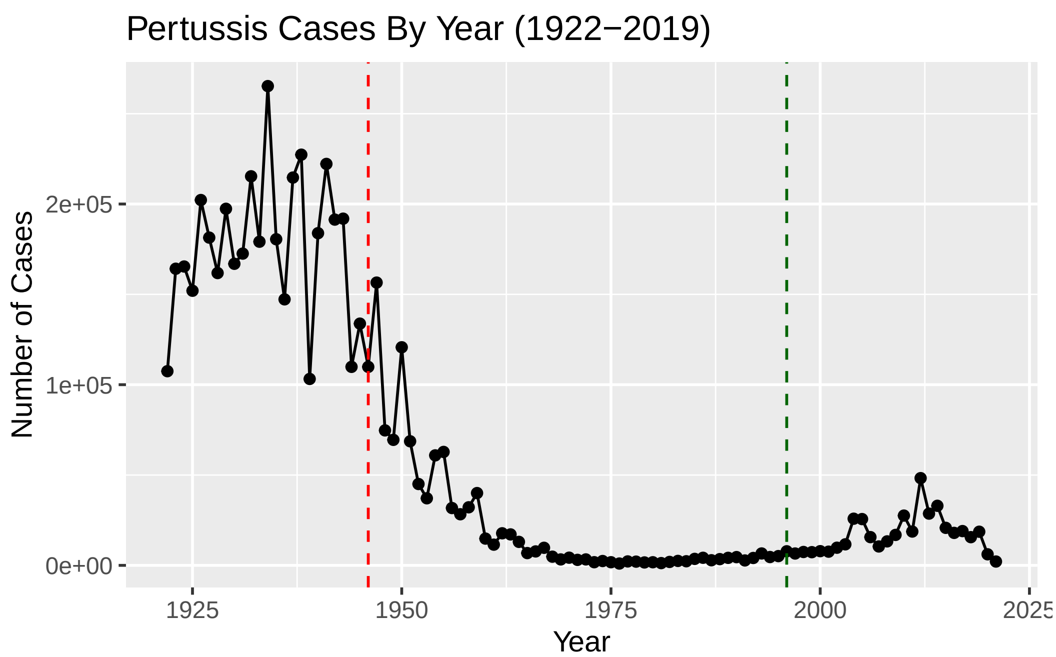
```r
ggplot(cdc) +
  aes(Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +
  labs(x = "Year", y = "Number of Cases", title = "Pertussis Cases By Year (1922-2019)")
```



A Tale of Two Vaccines (wP & aP)

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(Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +
  labs(x = "Year", y = "Number of Cases", title = "Pertussis Cases By Year (1922-2019)") +
  geom_vline(xintercept=1946, linetype = "dashed", color = "red") +
  geom_vline(xintercept=1996, linetype = "dashed", color = "darkgreen")
```



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

There was a slight increase in the number of cases, and a period where there was discrepancy in the number of cases. After the discrepancies, the cases dropped.

Exploring CMI-PB Data

```r
library(jsonlite)
```

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

```r
head(subject, 3)
```

```
  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
  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?

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

```

aP wP 
60 58 
```

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

```r
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…)?

```r
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
```

```r
library(lubridate)
```

```

Attaching package: 'lubridate'
```

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

    date, intersect, setdiff, union
```

```r
today()
```

```
[1] "2024-06-10"
```

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

```
Time difference of 8927 days
```

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

```
[1] 24.44079
```

> 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?

```r
today()
```

```
[1] "2024-06-10"
```

```r
today() - ymd(subject$year_of_birth)
```

```
Time differences in days
  [1] 14040 20615 15136 13310 12214 13310 15866 14405 10388 15501 14040 15501
 [13] 10022 11483 12944 13675 16232 10022 11118 15866 15136 14405 12214 11849
 [25] 13310 15136 10022 15501 10022 13310 12944 10022 12579 15136 12214 10022
 [37]  9657 10022 14405 11118 14405 10022  9657  9657 10022  9657 10388  9657
 [49] 10022 10022 10022  9657  9657 10022 10022 10022 10388 10022 10022 10022
 [61] 13675 11483 10753 11483 12579 17693 19154 19154 12579  9657  9657 12214
 [73] 10753 10753  9657  9657 13310 11483 13675 11849 11483  9657  9292 10022
 [85]  8927  9657  8927  8927 10022  9292  9657  8927 10388  9292  9657  8927
 [97] 14040 11483  9292  8561  7831  7831 11118 12944 11118 10388  9657 10753
[109] 12944 10022 10388 10388 10388 12579  8196  8927 11118  9657
```

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

```
  [1] 38.43943 56.44079 41.44011 36.44079 33.44011 36.44079 43.43874 39.43874
  [9] 28.44079 42.43943 38.43943 42.43943 27.43874 31.43874 35.43874 37.44011
 [17] 44.44079 27.43874 30.43943 43.43874 41.44011 39.43874 33.44011 32.44079
 [25] 36.44079 41.44011 27.43874 42.43943 27.43874 36.44079 35.43874 27.43874
 [33] 34.43943 41.44011 33.44011 27.43874 26.43943 27.43874 39.43874 30.43943
 [41] 39.43874 27.43874 26.43943 26.43943 27.43874 26.43943 28.44079 26.43943
 [49] 27.43874 27.43874 27.43874 26.43943 26.43943 27.43874 27.43874 27.43874
 [57] 28.44079 27.43874 27.43874 27.43874 37.44011 31.43874 29.44011 31.43874
 [65] 34.43943 48.44079 52.44079 52.44079 34.43943 26.43943 26.43943 33.44011
 [73] 29.44011 29.44011 26.43943 26.43943 36.44079 31.43874 37.44011 32.44079
 [81] 31.43874 26.43943 25.44011 27.43874 24.44079 26.43943 24.44079 24.44079
 [89] 27.43874 25.44011 26.43943 24.44079 28.44079 25.44011 26.43943 24.44079
 [97] 38.43943 31.43874 25.44011 23.43874 21.44011 21.44011 30.43943 35.43874
[105] 30.43943 28.44079 26.43943 29.44011 35.43874 27.43874 28.44079 28.44079
[113] 28.44079 34.43943 22.43943 24.44079 30.43943 26.43943
```

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

```
Time differences in days
  [1] 14040 20615 15136 13310 12214 13310 15866 14405 10388 15501 14040 15501
 [13] 10022 11483 12944 13675 16232 10022 11118 15866 15136 14405 12214 11849
 [25] 13310 15136 10022 15501 10022 13310 12944 10022 12579 15136 12214 10022
 [37]  9657 10022 14405 11118 14405 10022  9657  9657 10022  9657 10388  9657
 [49] 10022 10022 10022  9657  9657 10022 10022 10022 10388 10022 10022 10022
 [61] 13675 11483 10753 11483 12579 17693 19154 19154 12579  9657  9657 12214
 [73] 10753 10753  9657  9657 13310 11483 13675 11849 11483  9657  9292 10022
 [85]  8927  9657  8927  8927 10022  9292  9657  8927 10388  9292  9657  8927
 [97] 14040 11483  9292  8561  7831  7831 11118 12944 11118 10388  9657 10753
[109] 12944 10022 10388 10388 10388 12579  8196  8927 11118  9657
```

```r
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?

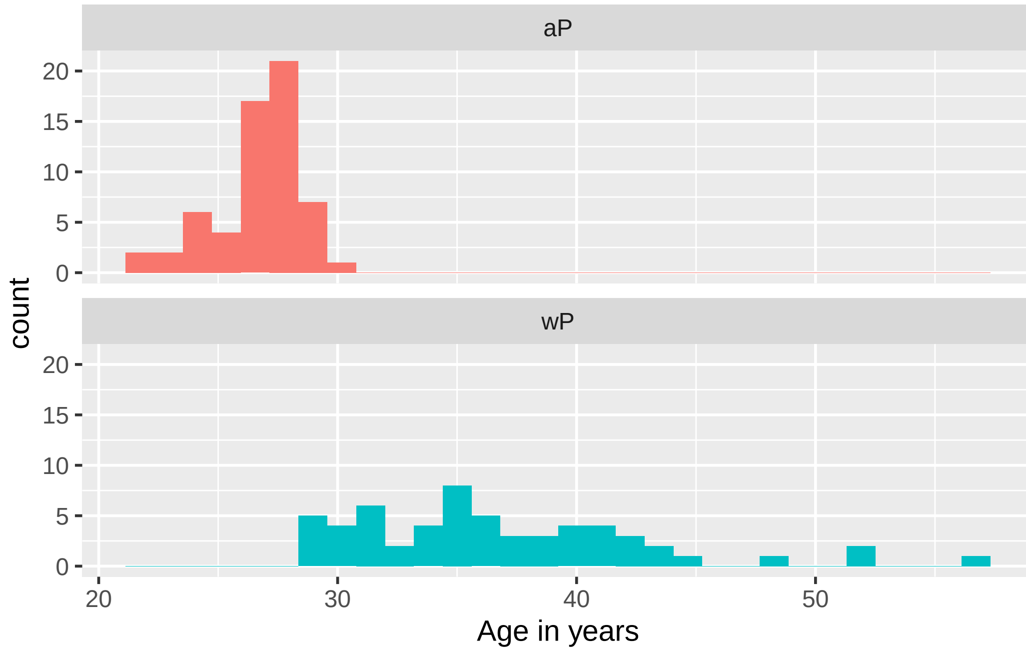
```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

```
[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`.




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

x$p.value
```

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

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

Joining Multiple Tables

```
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:

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

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

```r
dim(meta)
```

```
[1] 939  14
```

```r
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 14040 days
2 14040 days
3 14040 days
4 14040 days
5 14040 days
6 14040 days
```

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.

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

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

```
dim(abdata)
```

```
[1] 46906    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 4255 8983 8990 8990 8990 
```

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         7301 
```

Examine IgG Ab Titer Levels

```
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
```

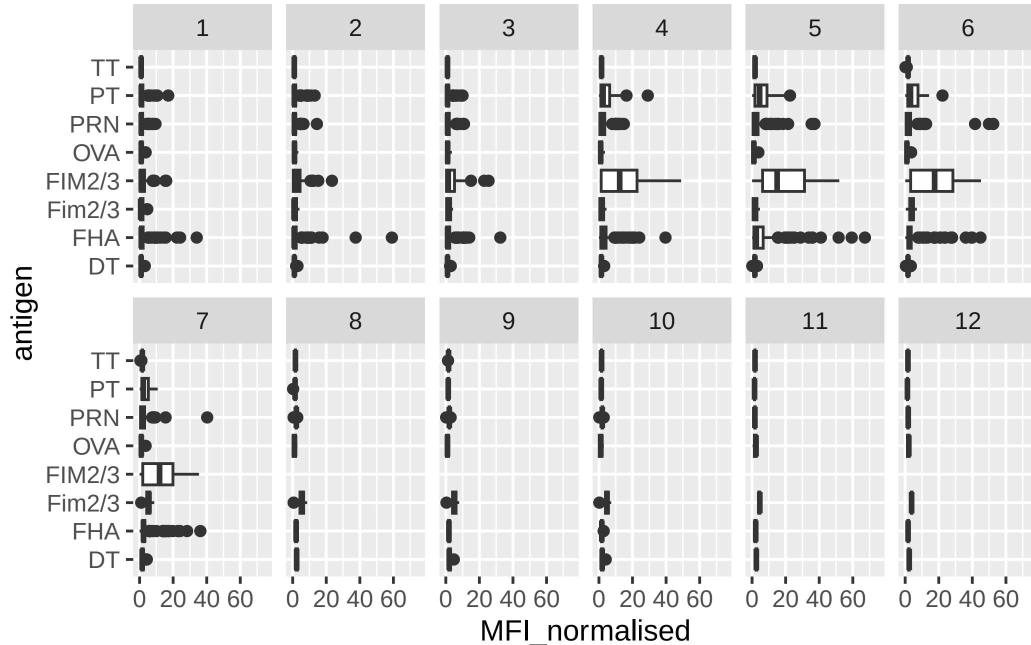
```
4          19     IgG                TRUE      PT    20.11607       1.096366
5          19     IgG                TRUE     PRN   976.67419       7.652635
6          19     IgG                TRUE     FHA    60.76626       1.096457
   unit lower_limit_of_detection subject_id actual_day_relative_to_boost
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 14040 days
2 14040 days
3 14040 days
4 15136 days
5 15136 days
6 15136 days
```

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() +
    xlim(0,75) +
  facet_wrap(vars(visit), nrow=2)
```

```
Warning: Removed 5 rows containing non-finite outside the scale range
```

```
(`stat_boxplot()`).
```

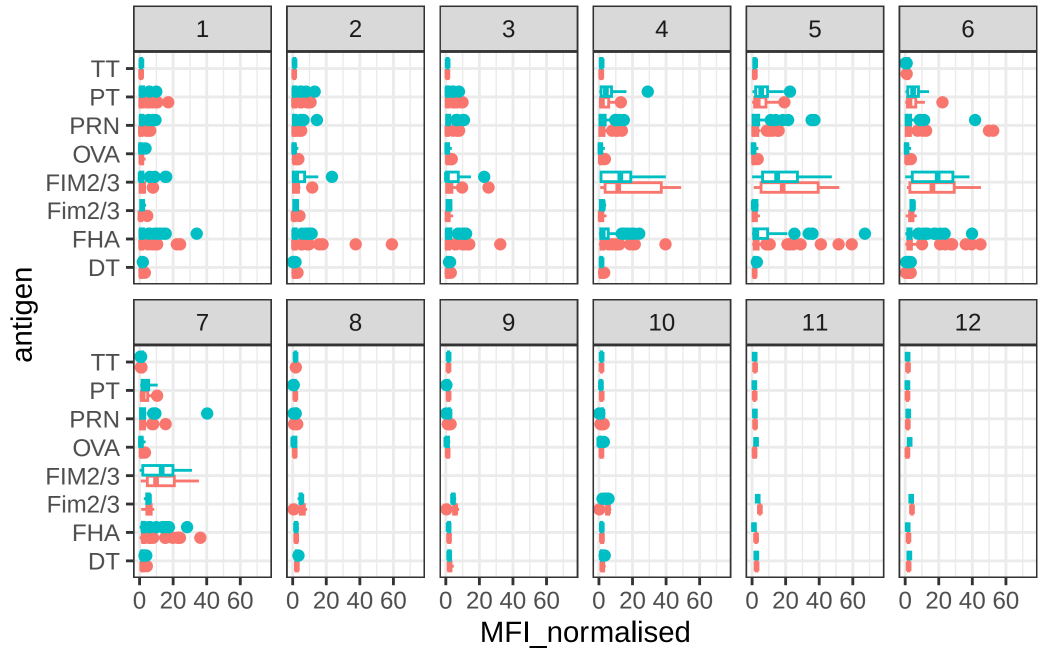


> Q14. What antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others?

FHA antigen shows the most difference.

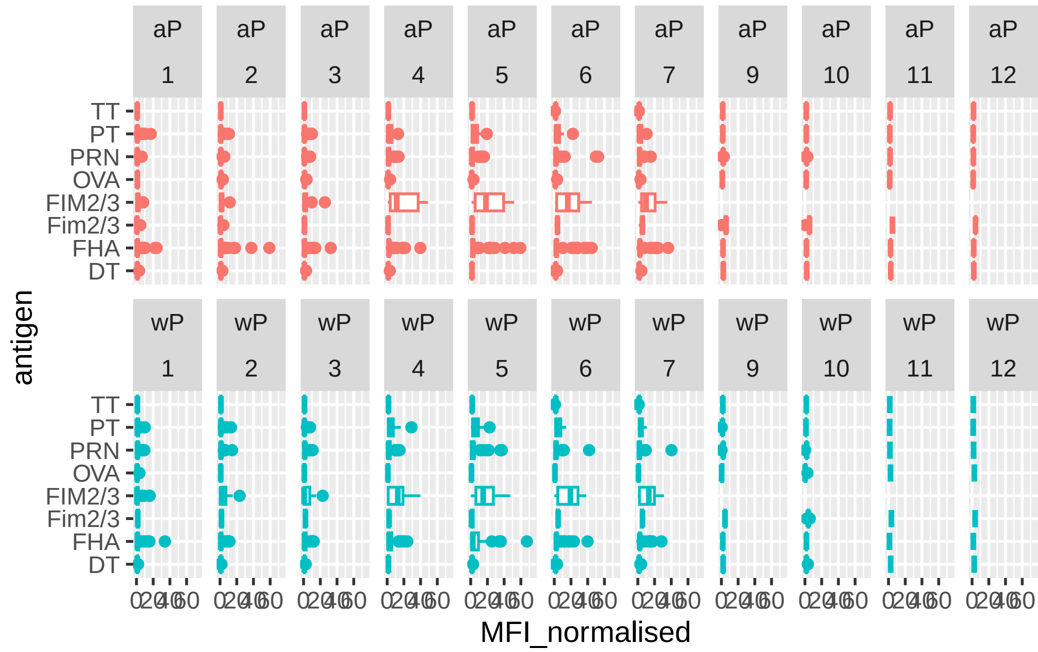
```r
ggplot(igg) +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  xlim(0,75) +
  theme_bw()
```

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



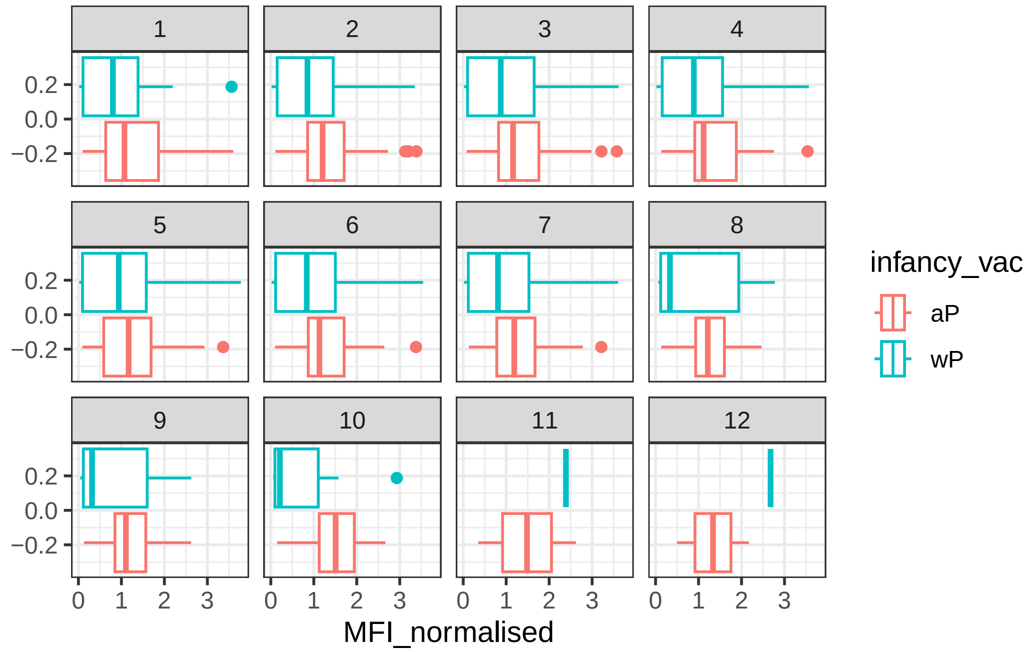

```
igg %>% filter(visit != 8) %>%
ggplot() +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  xlim(0,75) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```

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

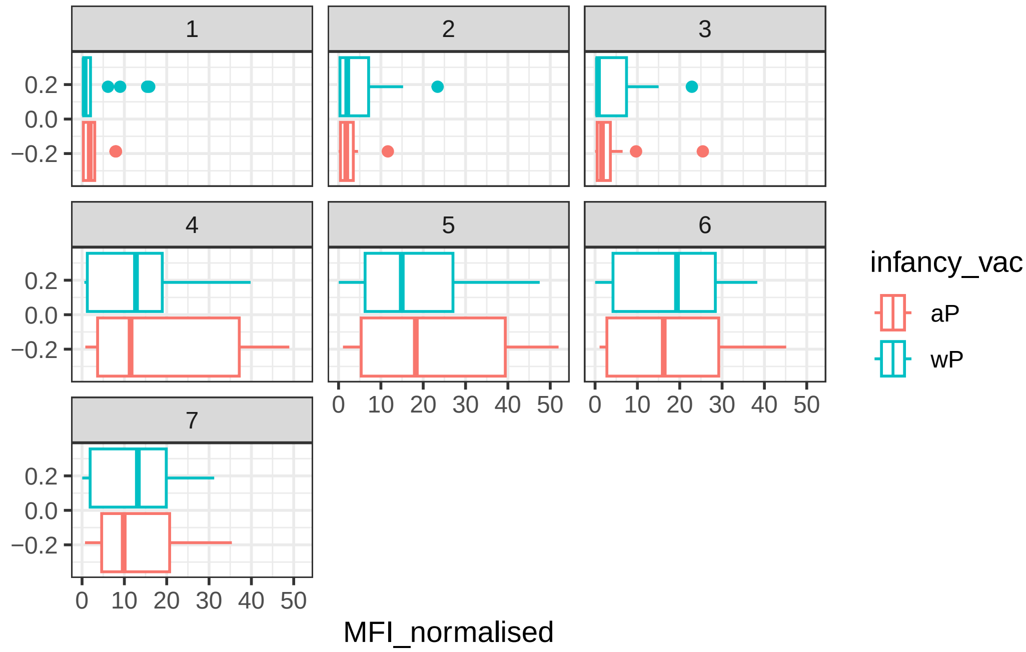


Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("OVA", that is not in our vaccines) and a clear antigen of interest ("PT", Pertussis Toxin, one of the key virulence factors produced by the bacterium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

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




```
filter(igg, antigen=="FIM2/3") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = TRUE) +
  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 are higher than OVA levels. PT data for aP and wP is quite similar, while OVA data for aP and wP is quite different.

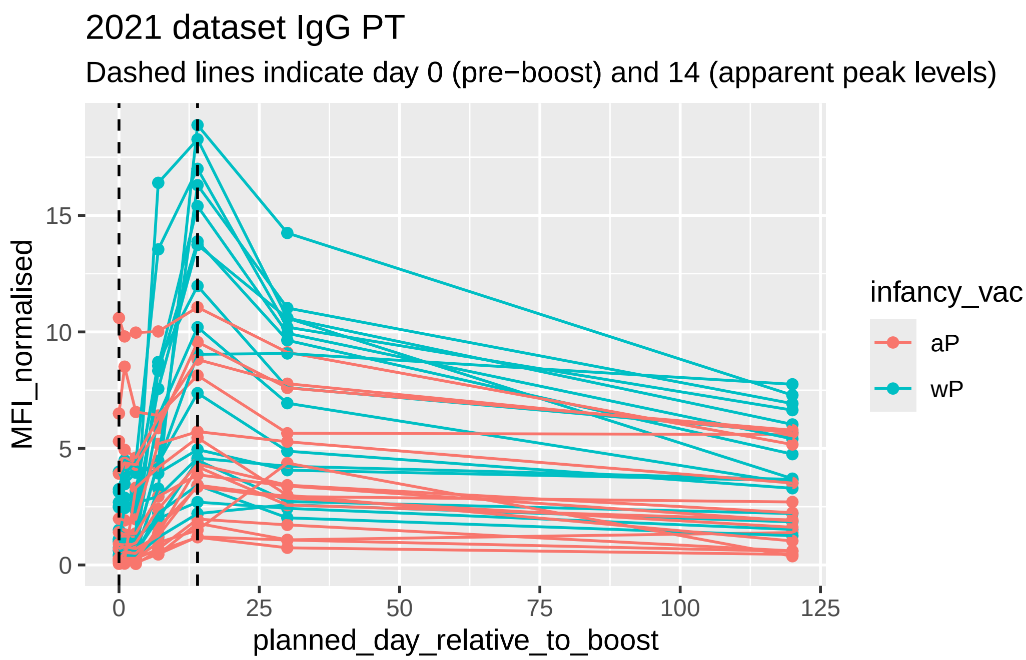
Q17. Do you see any clear difference in aP vs. wP responses?

In OVA, yes. In PT, no.

```
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") +
```

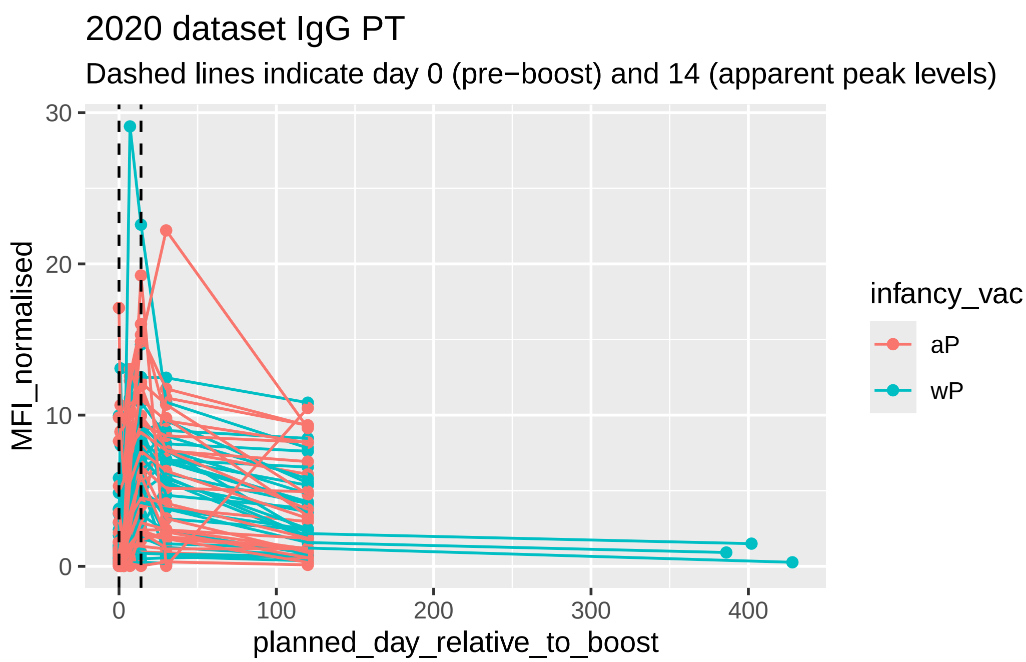
```
  labs(title="2021 dataset IgG PT",
       subtitle = "Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)")
```



Q18. Does this trend look similar for the 2020 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)")
```



Somewhat similar trend, however 2020 has some unique trends that do not appear in the 2021 data.

Obtaining CMI-PB RNASeq Data

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

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

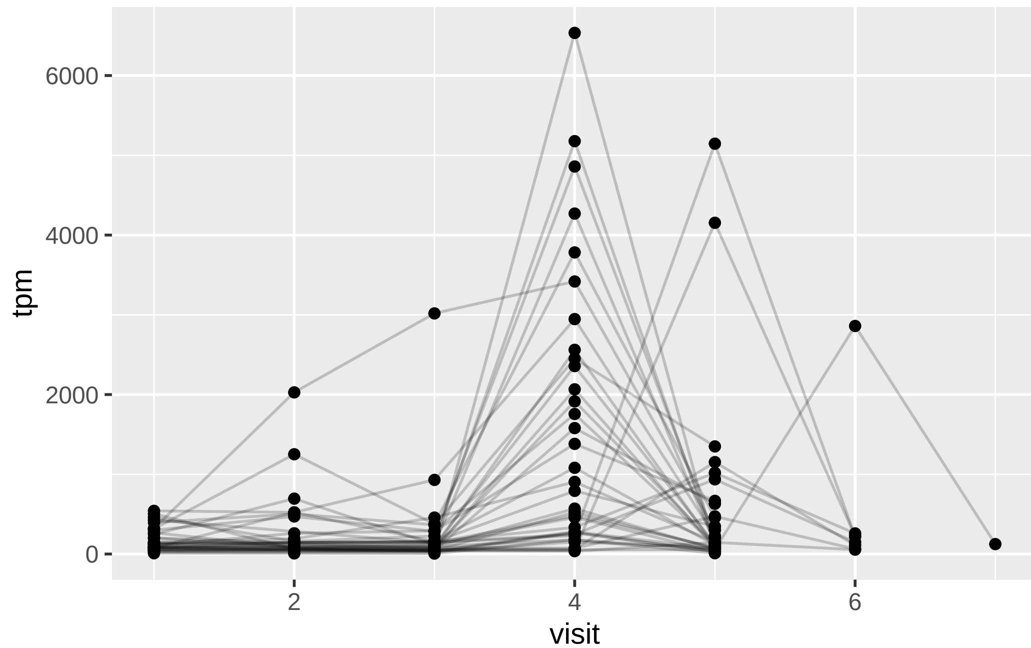
```
#meta <- inner_join(specimen, subject)
ssrna <- inner_join(rna, meta)
```

```
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)
```



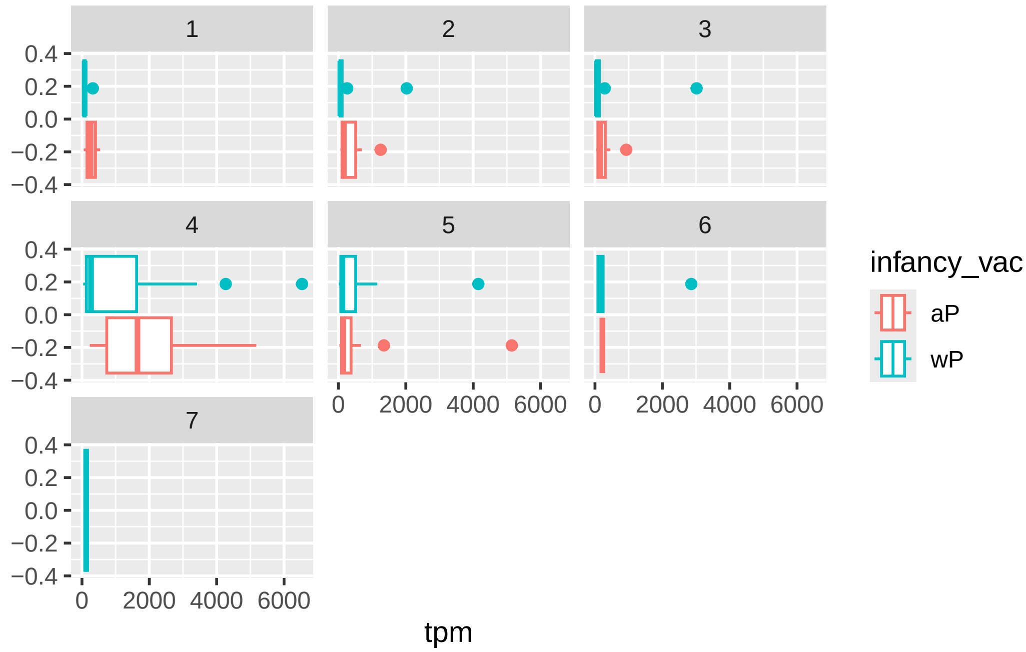
> Q20.: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

At visit 4, tpm is at its highest.

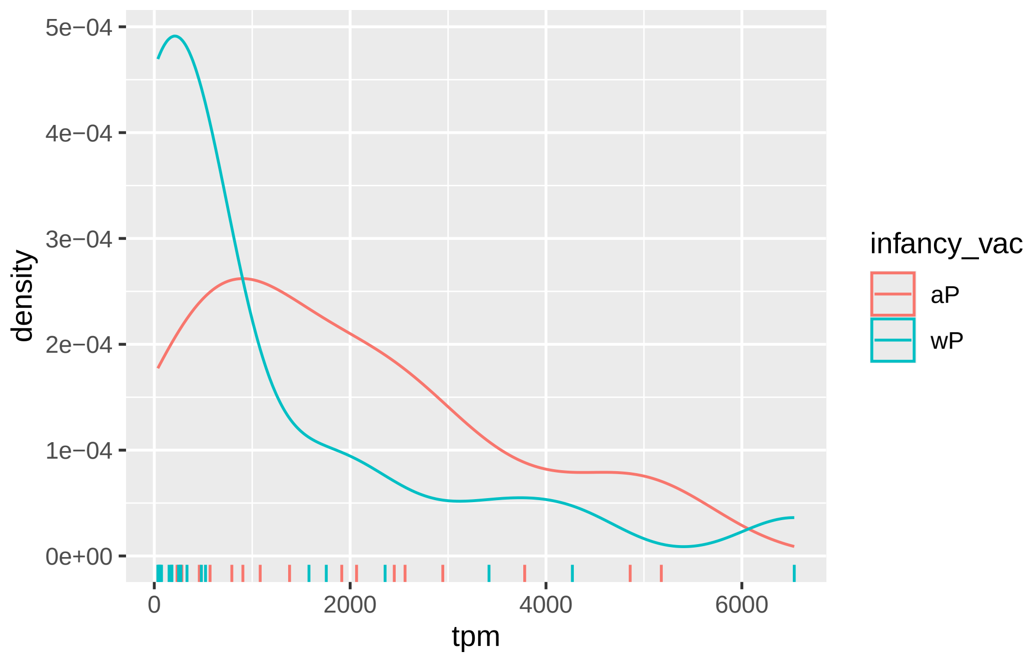
> Q21. Does this pattern in time match the trend of antibody titer data? If not, why not?

No. At visit 4, antigen levels were not high in the titer data.

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




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



density
5e−04
4e−04
3e−04
2e−04
1e−04
0e+00
0
2000
4000
6000
tpm
infancy_vac
aP
wP