Mini Project Pertussis

Morgan Farrell

3/9/2022

Import data from the CDC website on Pertussis. Use the package datapasta to get the information from the website to our R markdown.

```
#install.packages("datapasta")
```

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.

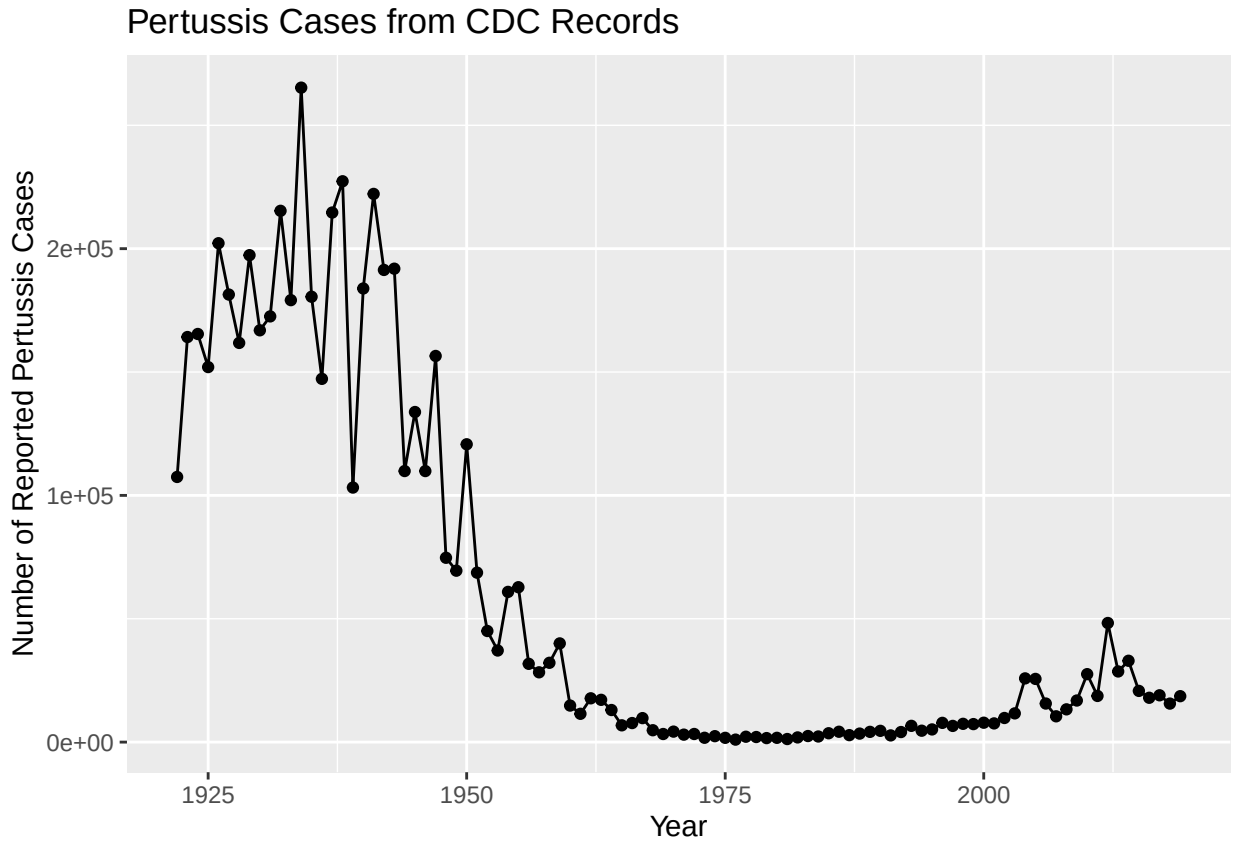
To get this data first create the object name and <- to add the data to it. Then copy the data from the website you are intersted in and then paste using the addins paste as data frame function

```
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)
      )
```

Graph this data to look at it

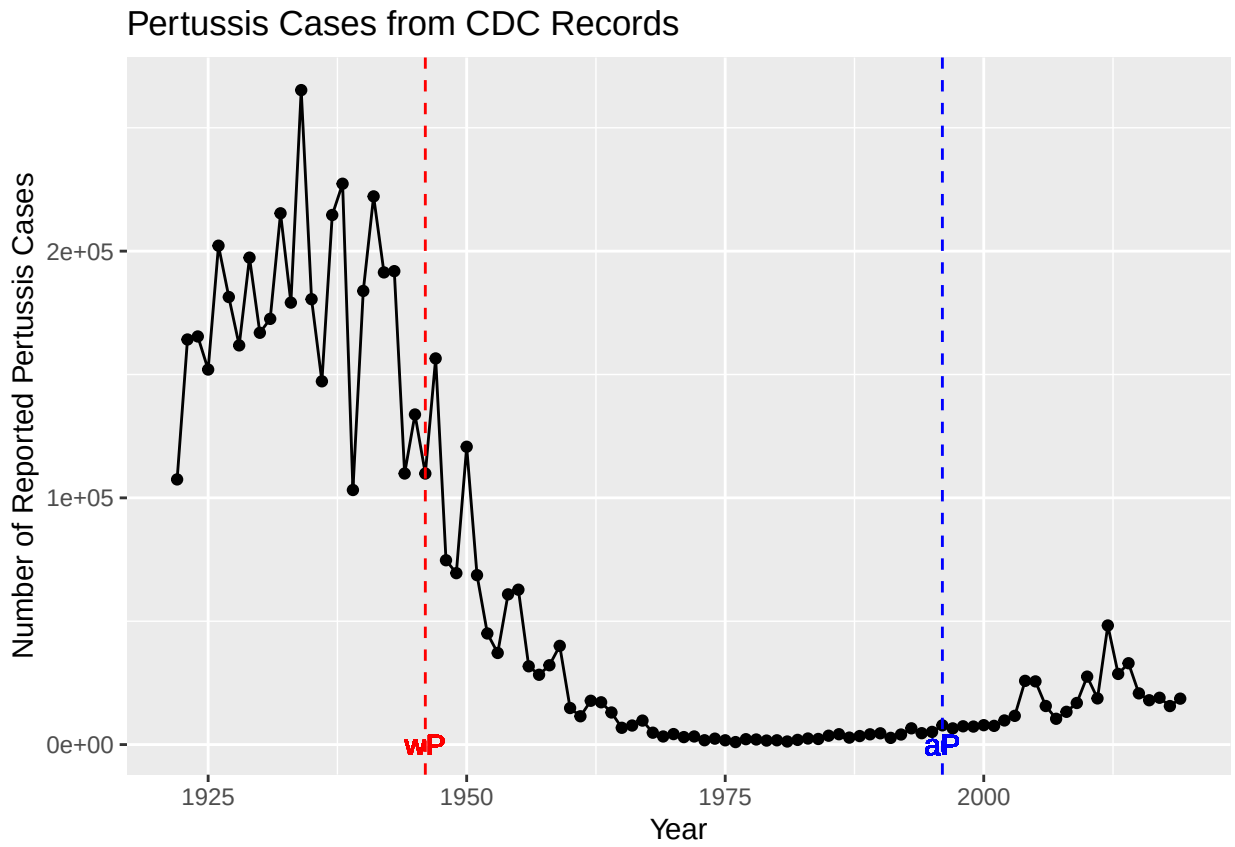
```
library(ggplot2)
ggplot(cdc, aes(x= Year, y= No..Reported.Pertussis.Cases)) +
    geom_point() +
    geom_line() +
    labs(x="Year", y="Number of Reported Pertussis Cases", title="Pertussis Cases from CDC Records")
```



There is a large number of cases in the 1920s-1950s and then the number of cases begins to drop rapidly by the 1960s/1970s. There are also large variations and this is because there are seasons of increased infection similar to flu season.

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(x= Year, y= No..Reported.Pertussis.Cases)) +
    geom_point() +
    geom_line() +
    geom_vline(xintercept= 1946, color= "red", linetype=2)+
    geom_vline(xintercept = 1996, color= "blue", linetype=2)+
    geom_text(x= 1946, y=0, label= "wP", color= "red")+
    geom_text(x= 1996, y=0, label= "aP", color= "blue")+
    labs(x="Year", y="Number of Reported Pertussis Cases", title="Pertussis Cases from CDC Records")
```



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

Many people were hesitant to vaccines which could have caused the increase in cases for the un-vaccinated people. There could have also been mutation of the Pertussis virus that introduced more cases.

One hypothesis is that there is an increase in cases in 10 year olds that were the first to receive this new aP vaccine. Potentially there is a waning immunity 10 years after the vaccination of the aP vaccine.

Exploring CMI-PD Data

We will be using the package jsonlite to read in the data

```
# Allows us to read, write and process JSON data
library(jsonlite)
```

Import data set from the CMI-PD website. By using the url we will keep updating our subject vector when new data is added.

```
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
## 2          2          wP         Female Not Hispanic or Latino White
## 3          3          wP         Female                 Unknown White
##   year_of_birth date_of_boost   study_name
## 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
```

Working with dates we will need teh lubridate package

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

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?

First I want to add a column that has the age of the subjects in years using tools from the lubridate packages

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

This is for the wP subjects

```
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
```

```
wP <- subject %>%
  filter(infancy_vac == "wP")

summary(wP$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   27.18   31.18   34.18   35.35   39.18   54.18
```

The mean age of wP vaccine subjects is 35.35

```
aP <- subject %>%
  filter(infancy_vac == "aP")

summary(aP$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.18   24.18   25.18   24.50   25.18   26.18
```

The mean age of aP vaccine subjects is 24.50. There is a 10 year difference between the average ages.

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

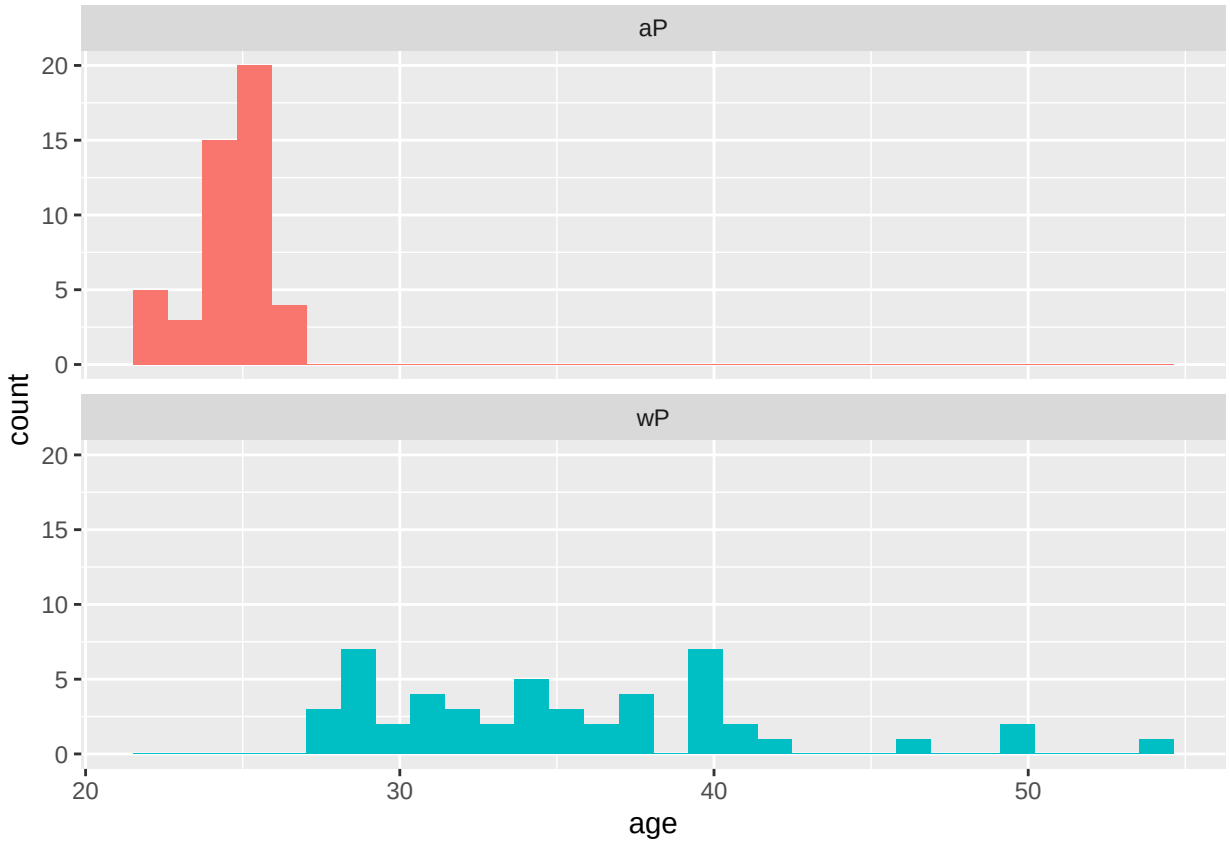
I will add a column to the dataset age_boost that will have the age each subject was when they were boost

```
subject$age_boost <- time_length(ymd(subject$date_of_boost)- ymd(subject$year_of_birth), "years")
```

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

```
ggplot(subject) +
  aes(age,
      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 first aP plot shows that the majority of the participants are younger than 30. While the wP participants are of a larger age range and mostly older than 30. This could make it difficult to compare both groups.

Joining multiple tables from the CMI-PD database

Import the new datasets

```
# Complete the API URLs...
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. 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:

We are using the inner join method over the full join method

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

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

```
dim(meta)
```

```
## [1] 729  15
```

```
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   study_name
## 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_boost
## 1 36.18344  30.69678
## 2 36.18344  30.69678
## 3 36.18344  30.69678
## 4 36.18344  30.69678
## 5 36.18344  30.69678
## 6 36.18344  30.69678
```

Since there are multiple specimens from each subject ID the number of rows jumps up to 729.

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, by = "specimen_id"
```

```
dim(abdata)
```

```
## [1] 32675    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 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
```

Visit 8 has significantly lower number of specimens than all the others, so it may not be good to include it in the dataset.

Examine the IgG1 Ab titer levels

subset the ig1 into its own vector for further analysis

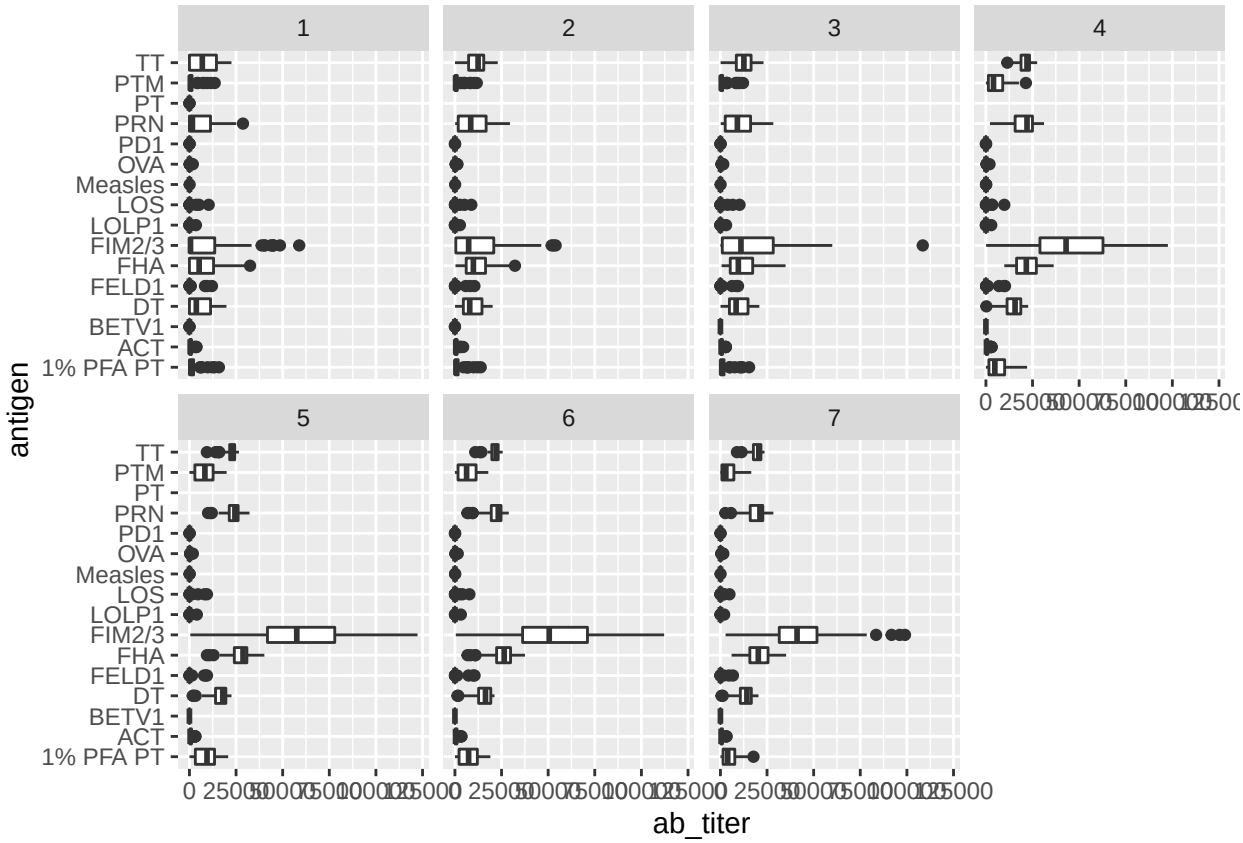
```
ig1 <- abdata %>% filter(isotype == "IgG1", visit!=8)
head(ig1)
```

```
##   specimen_id isotype is_antigen_specific antigen   ab_titer  unit
## 1           1    IgG1                TRUE     ACT 274.355068 IU/ML
## 2           1    IgG1                TRUE     LOS  10.974026 IU/ML
## 3           1    IgG1                TRUE   FELD1   1.448796 IU/ML
## 4           1    IgG1                TRUE   BETV1   0.100000 IU/ML
## 5           1    IgG1                TRUE   LOLP1   0.100000 IU/ML
## 6           1    IgG1                TRUE Measles  36.277417 IU/ML
##   lower_limit_of_detection subject_id actual_day_relative_to_boost
## 1                 3.848750          1                           -3
## 2                 4.357917          1                           -3
## 3                 2.699944          1                           -3
## 4                 1.734784          1                           -3
## 5                 2.550606          1                           -3
## 6                 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   study_name
## 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_boost
## 1 36.18344  30.69678
## 2 36.18344  30.69678
## 3 36.18344  30.69678
## 4 36.18344  30.69678
## 5 36.18344  30.69678
## 6 36.18344  30.69678
```

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

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

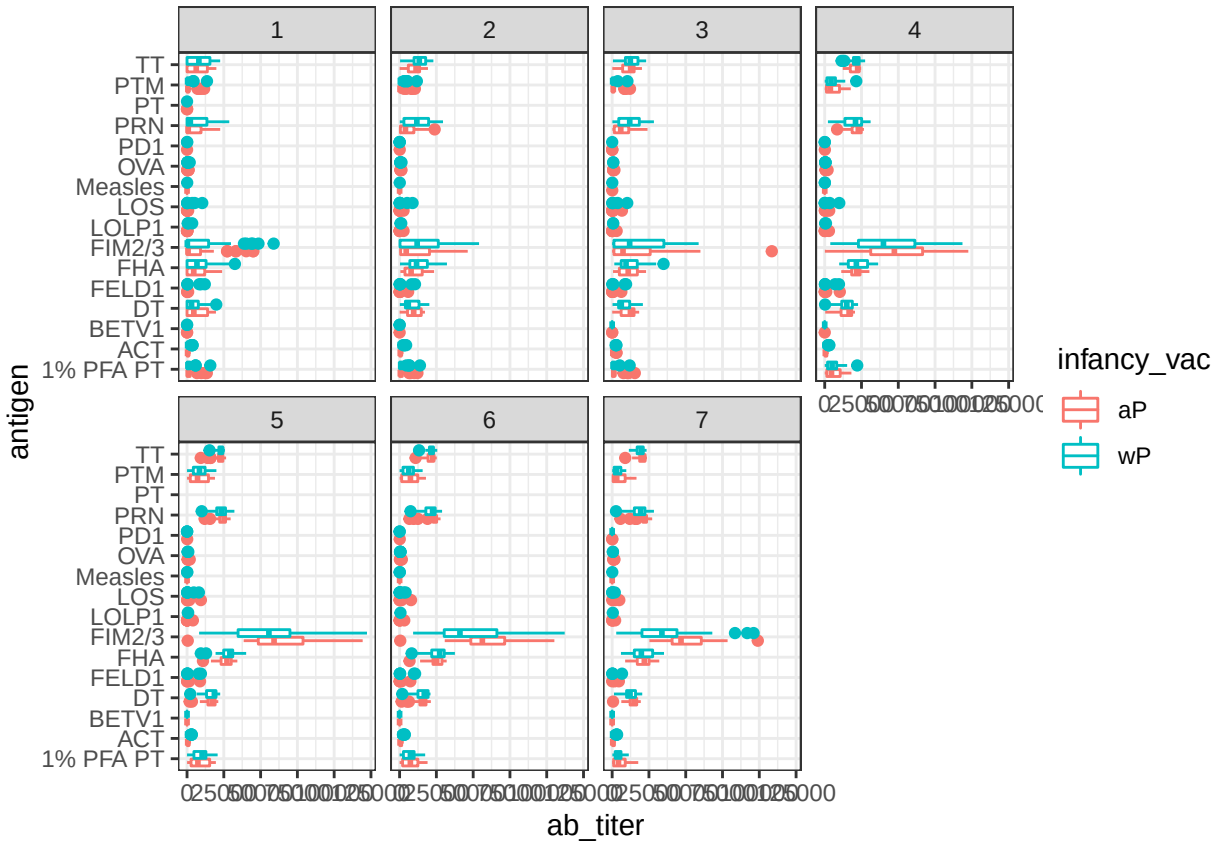


Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

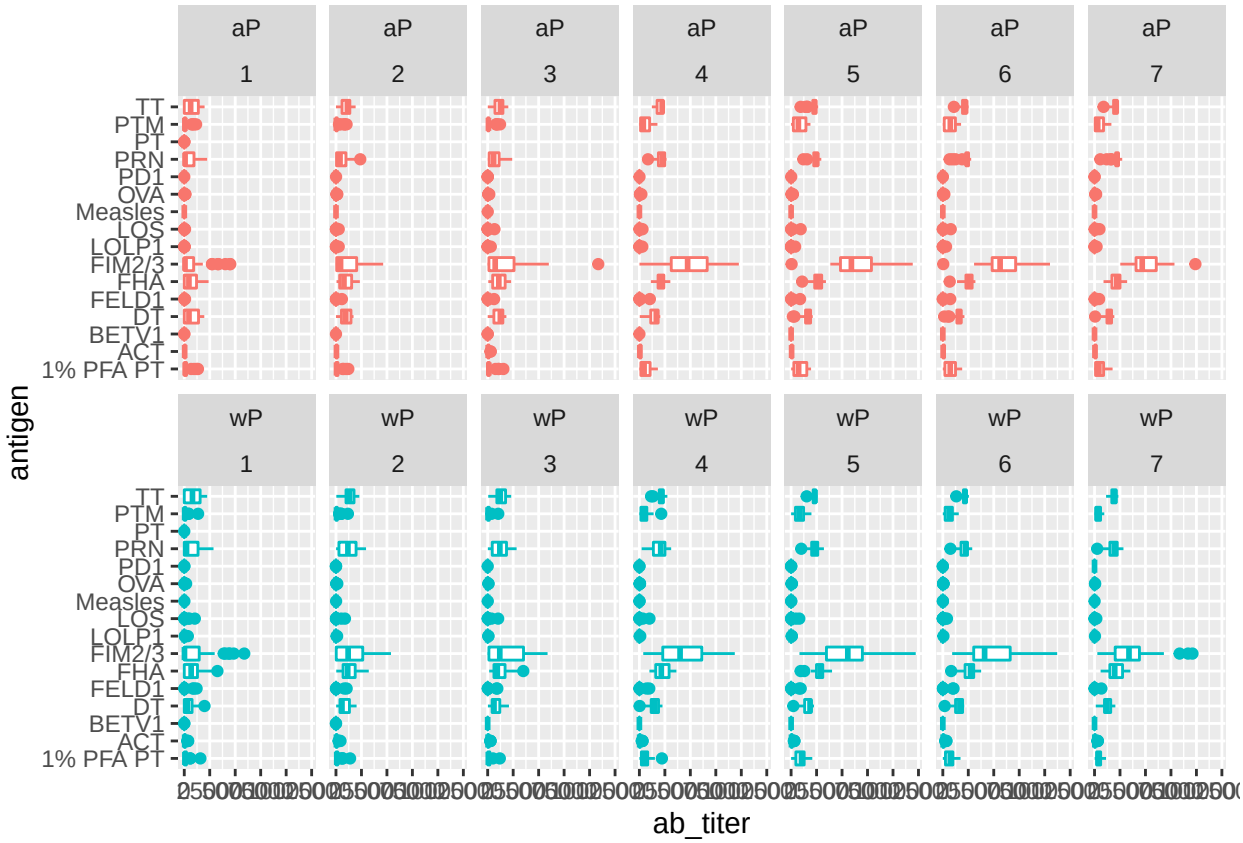
The antigens are the things being recognized by the antibody. FIM2/3 is really high for Ig1 antibody titer level compared to all others. Other notable ones are FHA and PRN are slightly higher as well. The rest are staying relatively low. Prn is higher because it is a protein that found in the vaccine, since it is a protein associated with Pertussis. TT is tetanus toxin which goes up slightly. Measles stays zero over time which makes sense since we are not testing for measles immunity. FIM2/3 goes up because it is part of the vaccine as well as being an adhesion molecule of the Bordatella pertussis virus itself.

Graphs to look at infancy vaccine type

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



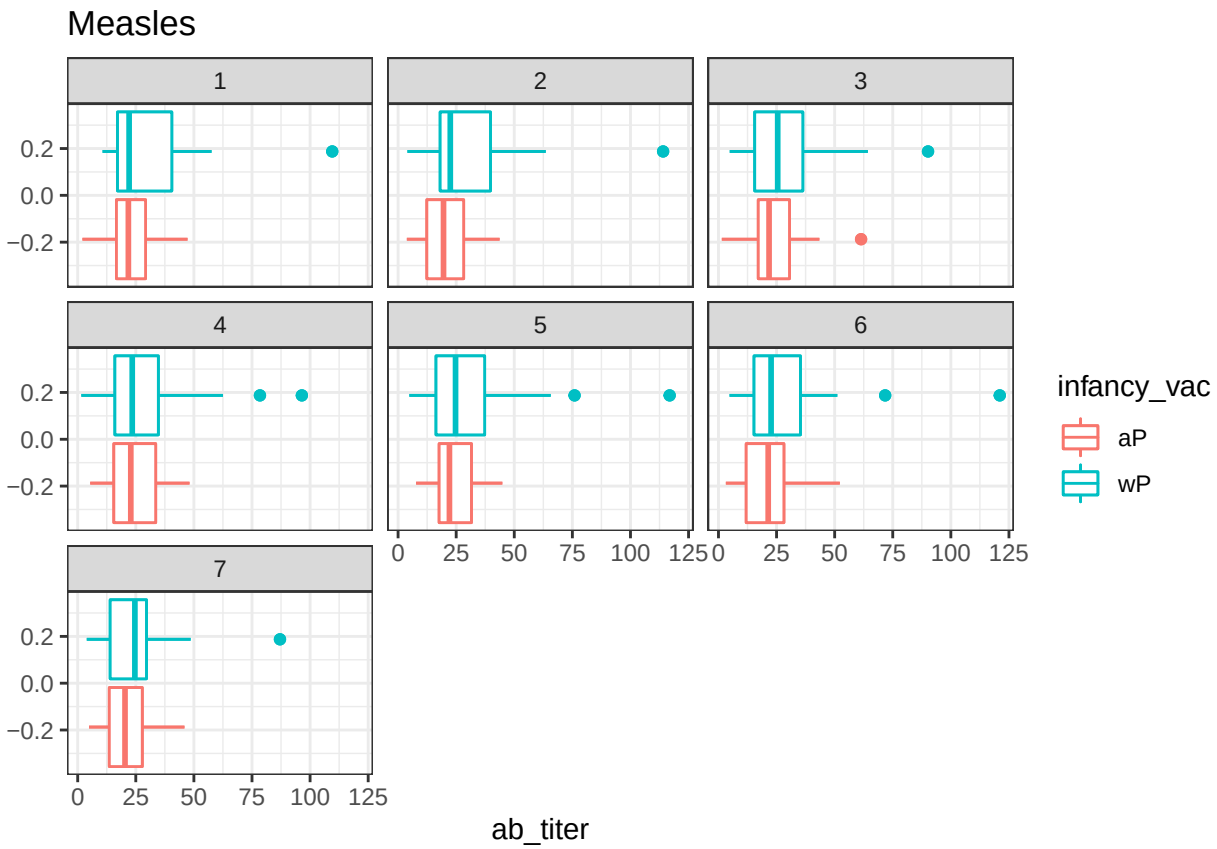

```
ggplot(ig1) +
  aes(ab_titer, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```



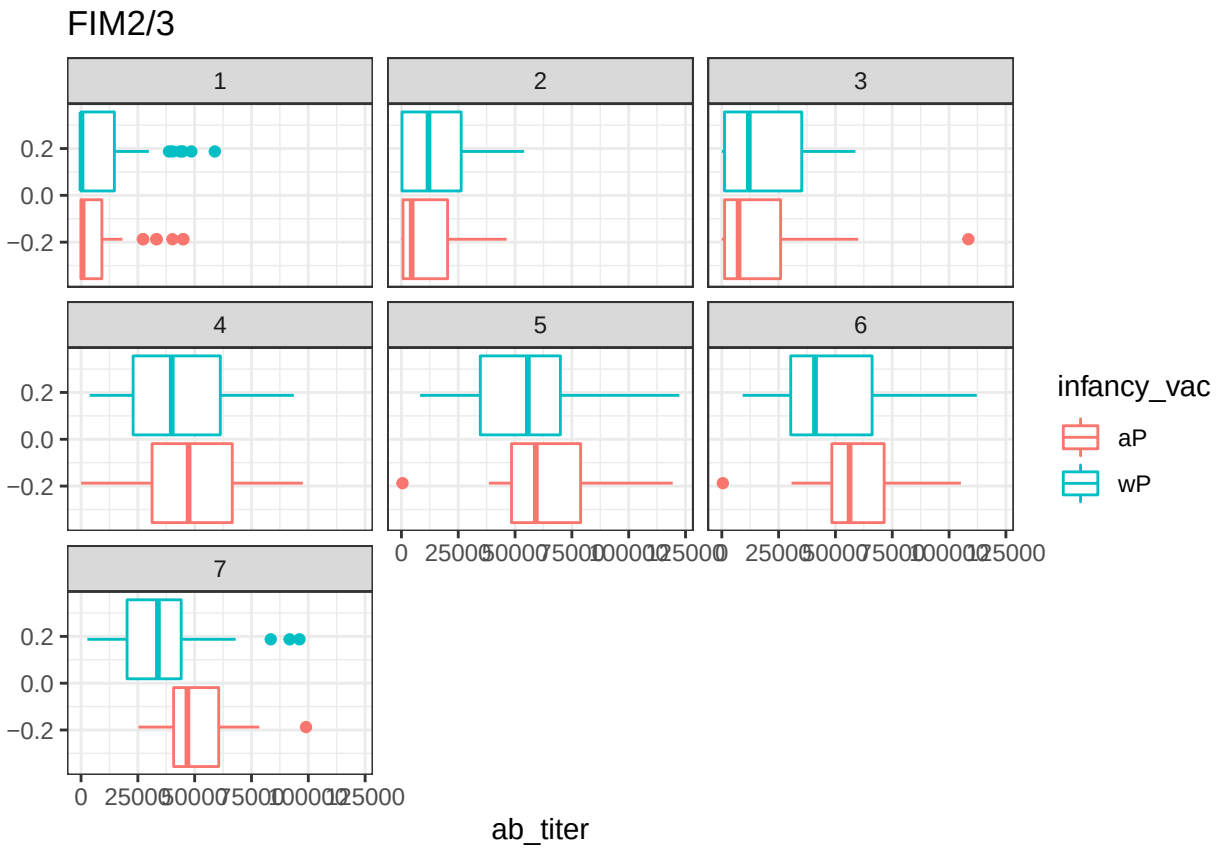
I think I prefer the second one, it is easier to read than the overlapping one before.

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(ab_titer, col=infancy_vac) +
  geom_boxplot() +
  facet_wrap(vars(visit)) +
  geom_boxplot(show.legend = T) +
  theme_bw()+
  labs(title= "Measles")
```




```
filter(ig1, antigen== "FIM2/3") %>%
  ggplot() +
  aes(ab_titer, col=infancy_vac) +
  geom_boxplot(show.legend = T) +
  facet_wrap(vars(visit)) +
  theme_bw() +
  labs(title= "FIM2/3")
```



Q16. What do you notice about these two antigens time courses and the FIM2/3 data in particular?

Measles antigen stays low over the course of the visits compared to FIM2/3 which increases a lot. FIM2/3 looks as though it peaks at visit 5 and then goes down slightly in visits 6 and 7.

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

The aP response was slightly higher than than the wP, but the error bars look like they overlap quite a bit so it doesn't look like there is a true difference.

RNA seq Data

```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.7"

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

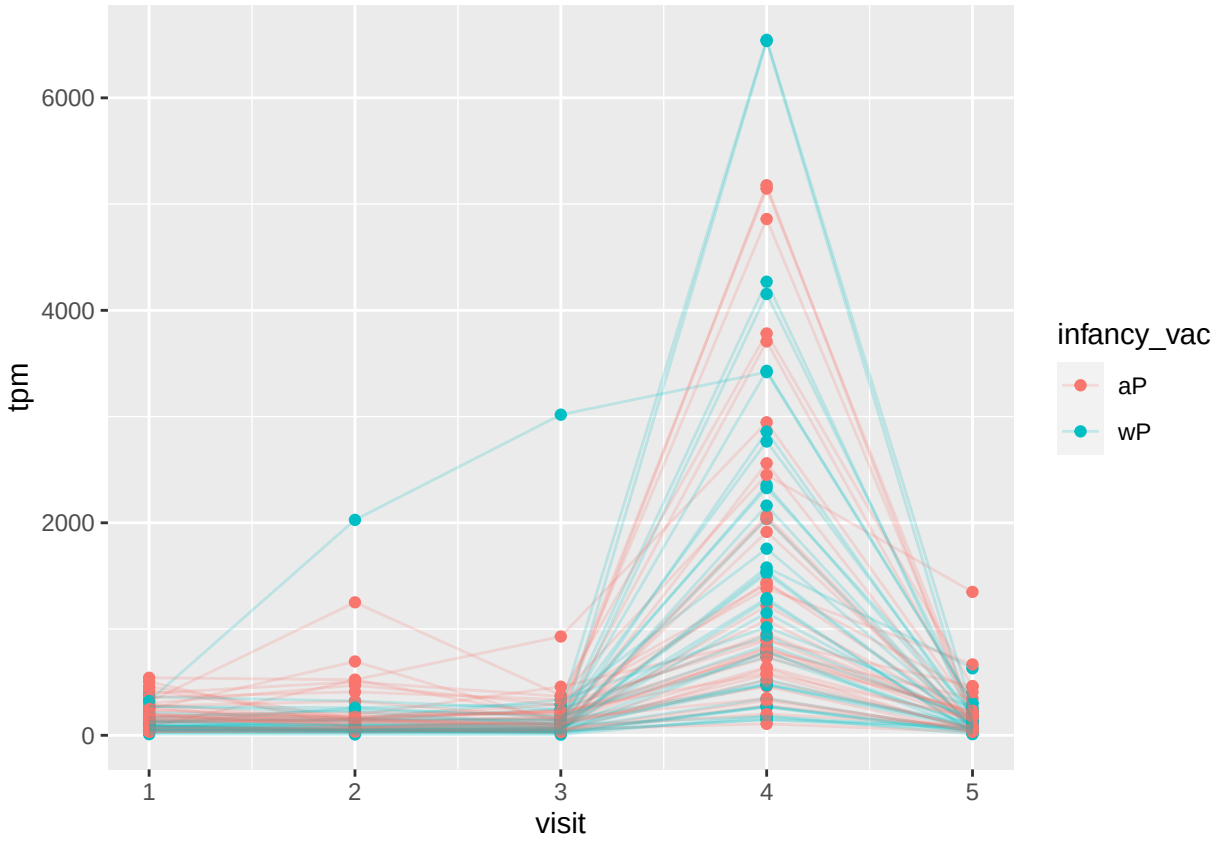
```
ssrna <- inner_join(rna, meta)
```

```
## Joining, 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).

Group by subject id so that the lines will be drawn properly between the points.

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



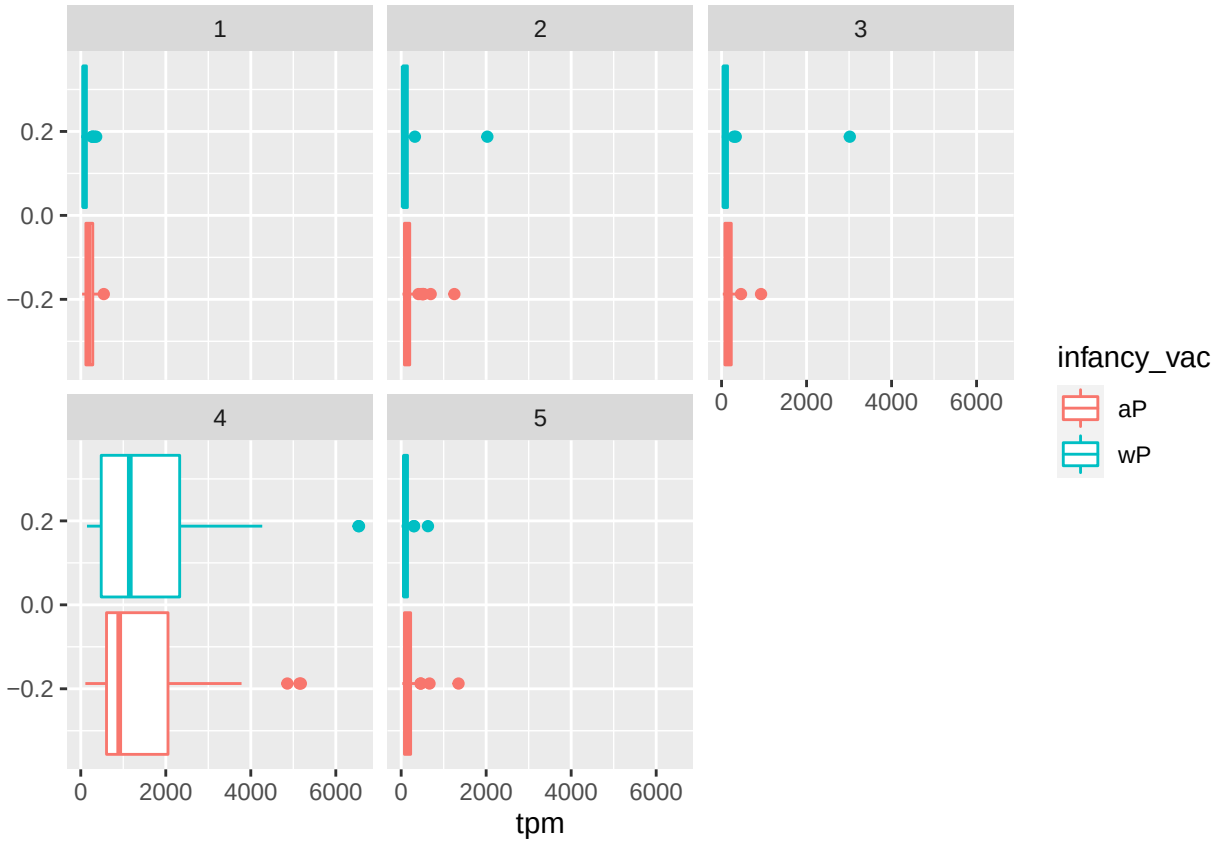
Q19.: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

The expression of this gene is highest at visit 4 and relatively low if not zero at all other visits.

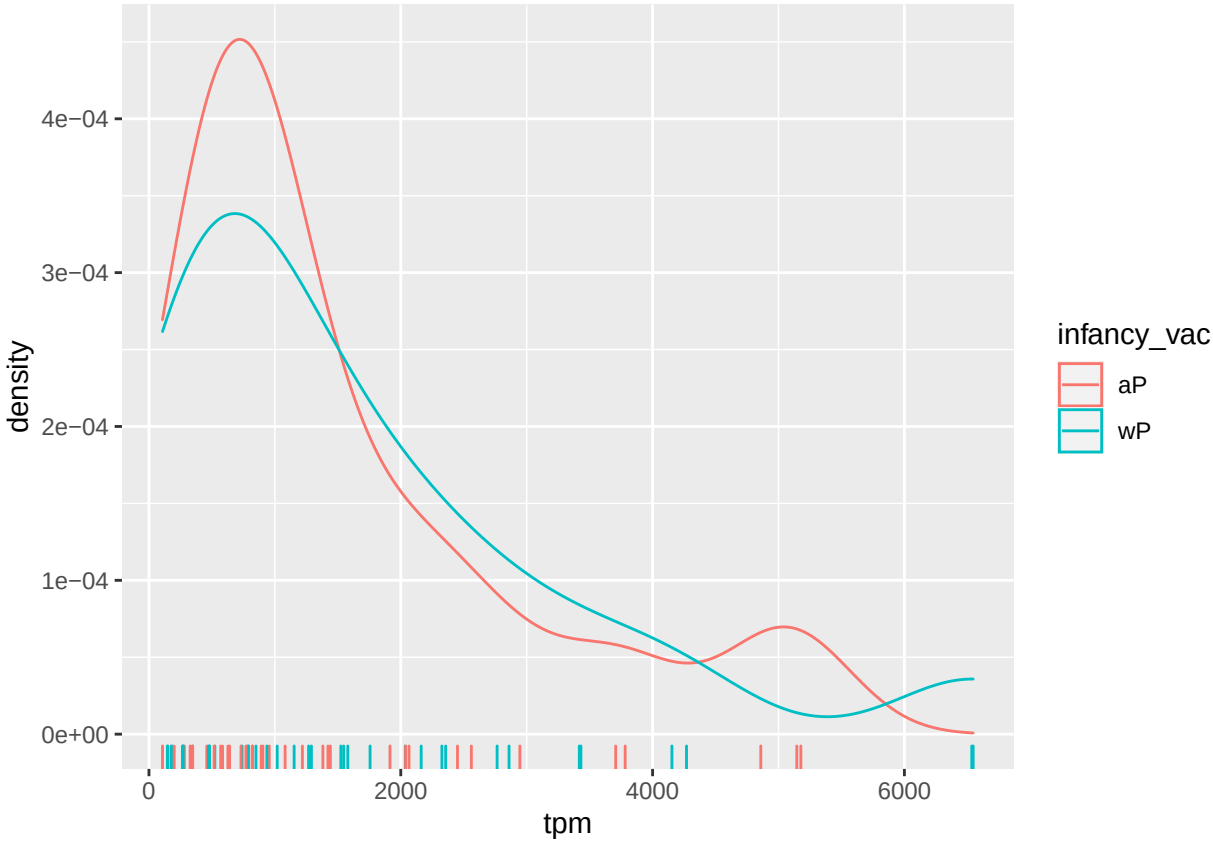
Q20. Does this pattern in time match the trend of antibody titer data? If not, why not?

It does slightly, I thought that some of the antigens peaked at visit 5 from the antibody data but that is very close to the expression levels peaking at visit 4. It might have a delay for the expression to peak at visit 4, but then the detection of antibodies wouldn't peak until a little later at visit 5 once these levels have increased their production. Additionally, antibodies will persist in the system for a long time and the cells will continue to make them. This explains why the antibody levels stay high in visits 5-7.

```
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()
```


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