Class 15 Pertussis Mini Project

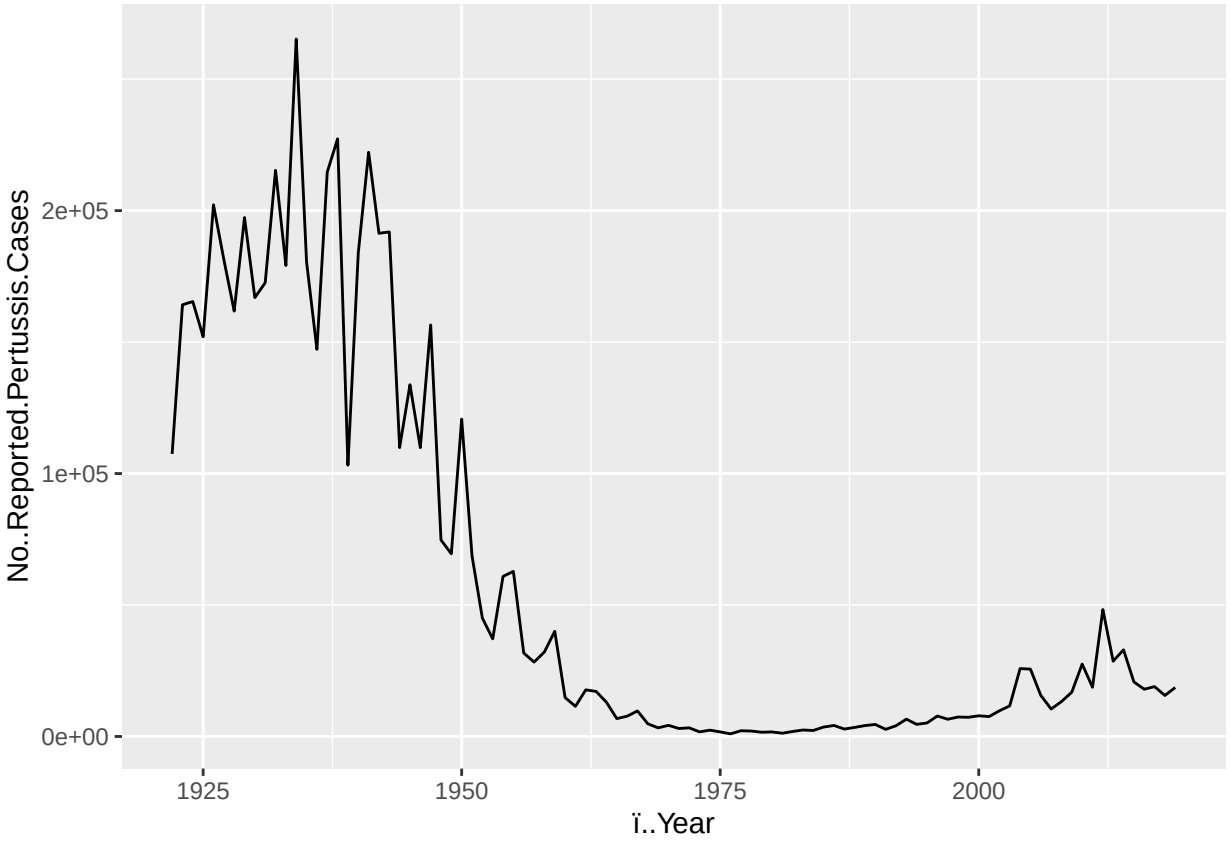
Cindy Tran

3/8/2022

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.

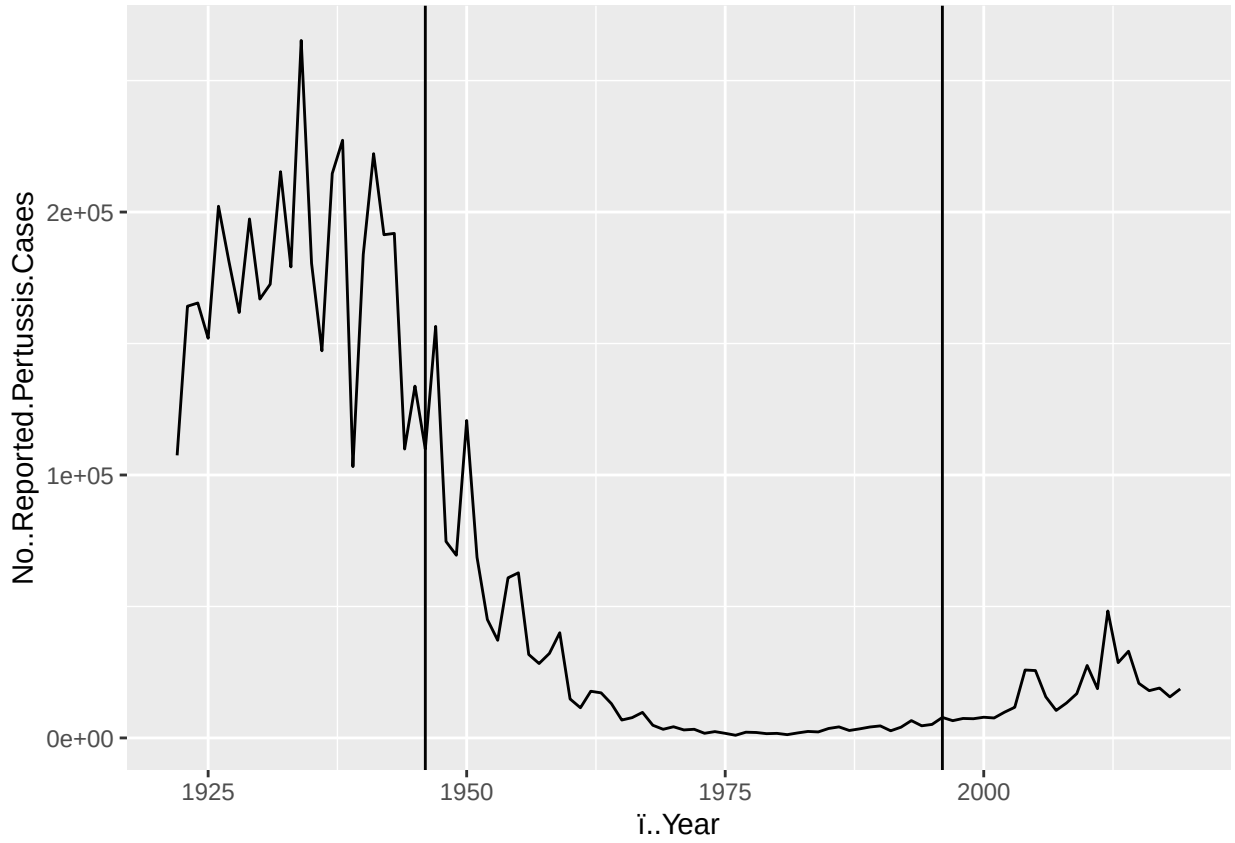
```
cdc <- data.frame(read.csv("BIMM143class15.csv"))

library(ggplot2)

ggplot(cdc) +
  aes(ï..Year, No..Reported.Pertussis.Cases) +
  geom_line()
```



2. A Tale of 2 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(ï..Year, No..Reported.Pertussis.Cases) +
  geom_line() +
  geom_vline(xintercept = 1946) +
  geom_vline(xintercept = 1996)
```



I noticed that cases dramatically dropped with the introduction of wP and increased slightly after introduction of aP

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

After the introduction of the aP vaccine, cases of pertussis rose slightly. This is poosibly due to aP not being as effective as wP against pertussis infection.

3. Exploring CMI-PB Data

```
# Allows us to read, write and process JSON data
library(jsonlite)
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 may 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$biological_sex, subject$race)
```

```
##
##          American Indian/Alaska Native Asian Black or African American
##   Female                             0    18                         2
##   Male                               1     9                         0
##
##          More Than One Race Native Hawaiian or Other Pacific Islander
##   Female                  8                                         1
##   Male                    2                                         1
##
##          Unknown or Not Reported White
##   Female                       10    27
##   Male                          4    13
```

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?

```
library(lubridate)
```

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

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

```
# Use todays date to calculate age in days
subject$age <- time_length(today() - ymd(subject$year_of_birth), "years")

#install.packages("https://cran.r-project.org/src/contrib/Archive/rlang/rlang_1.0.1.tar.gz", repo = NUL

library(rlang)
```

```
##
## Attaching package: 'rlang'
```

```
## The following objects are masked from 'package:jsonlite':
##
##     flatten, unbox
```

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

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
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
```

```
# wP
wp <- subject %>% filter(infancy_vac == "wP")
summary(wp$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   27.18   31.18   34.18   35.34   39.18   54.18
```

Yes, the wP patients are older.

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

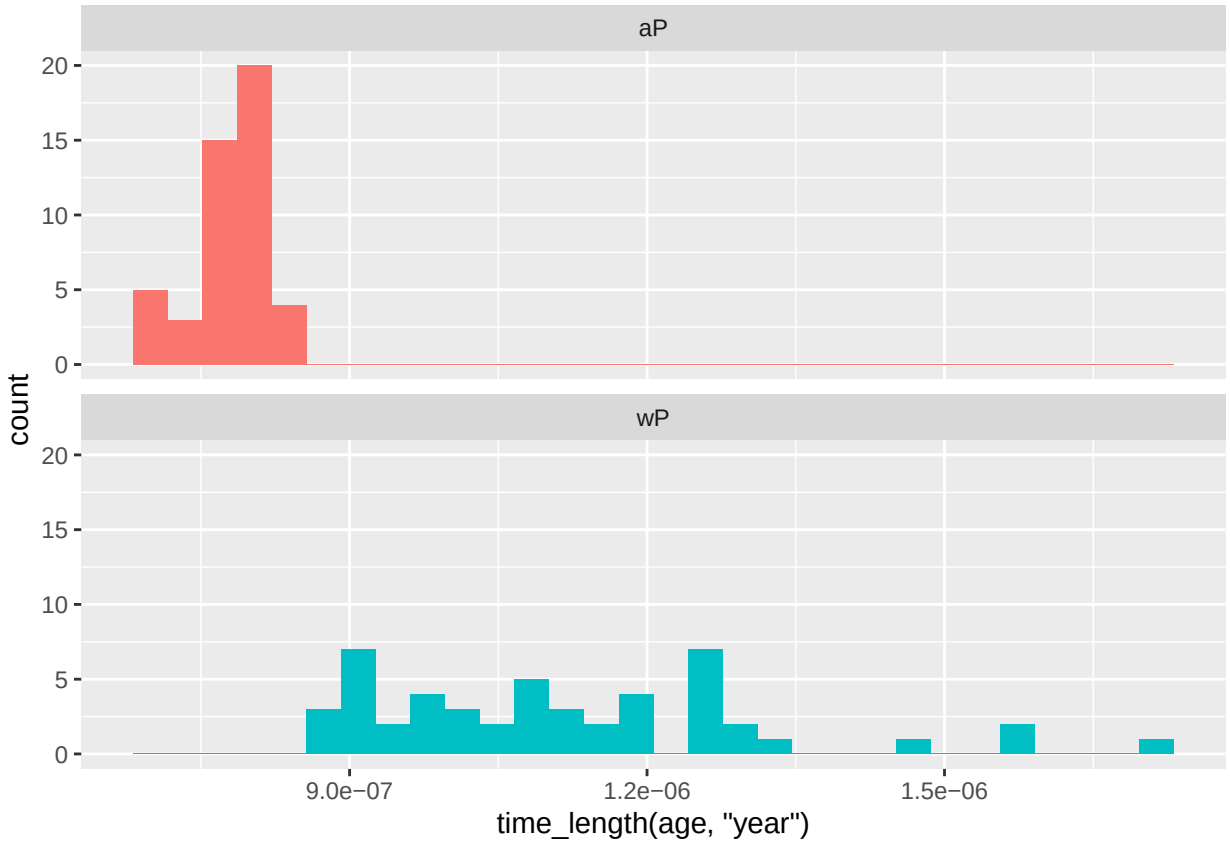
```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

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

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

```
library(ggplot2)
ggplot(subject) +
  aes(time_length(age, "year"),
      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`.
```



```
# 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:

```
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   study_name     age
## 1 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.1807
## 2 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.1807
## 3 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.1807
## 4 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.1807
## 5 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.1807
## 6 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.1807
```

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    20
```

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 much less specimens.

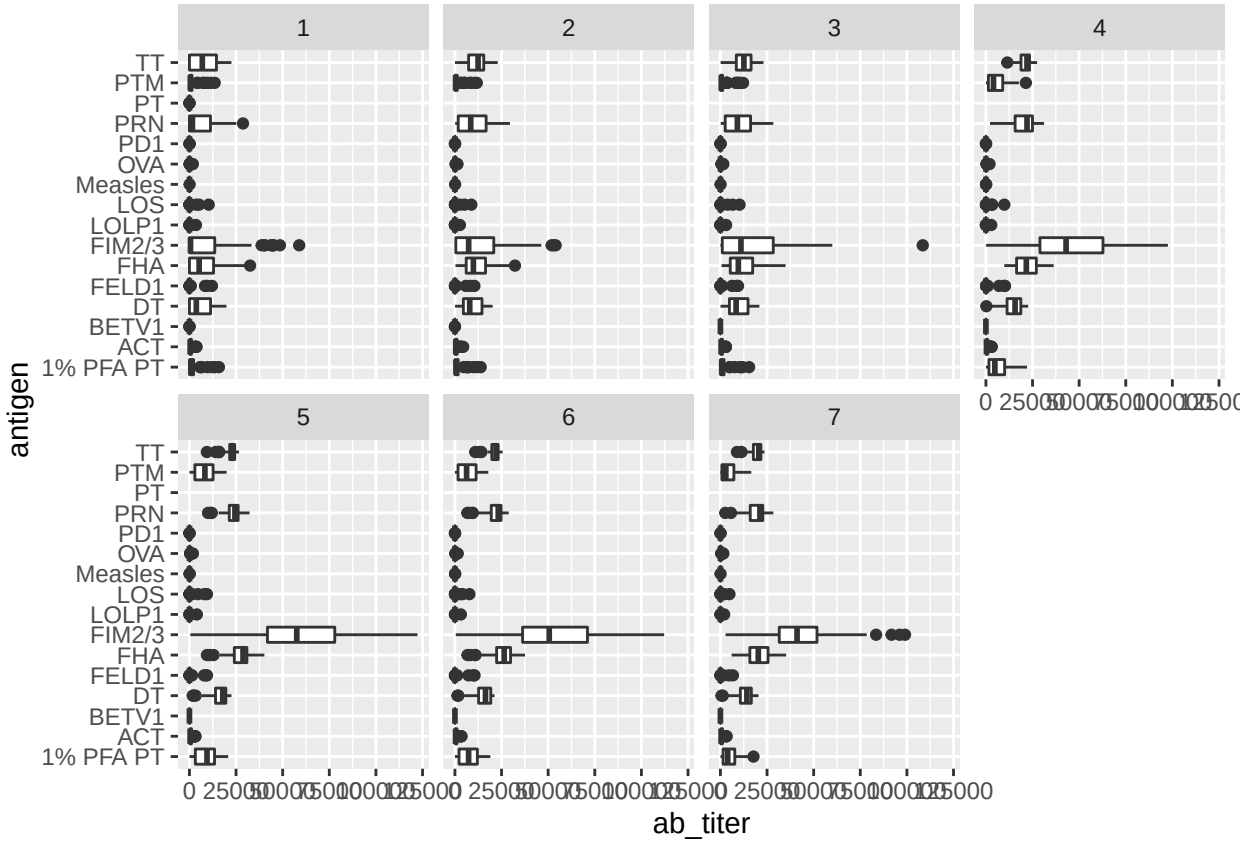
4. Examine IgG1 Ab Titer Levels

```
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     age
## 1 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.1807
## 2 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.1807
## 3 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.1807
## 4 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.1807
## 5 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.1807
## 6 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.1807
```

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

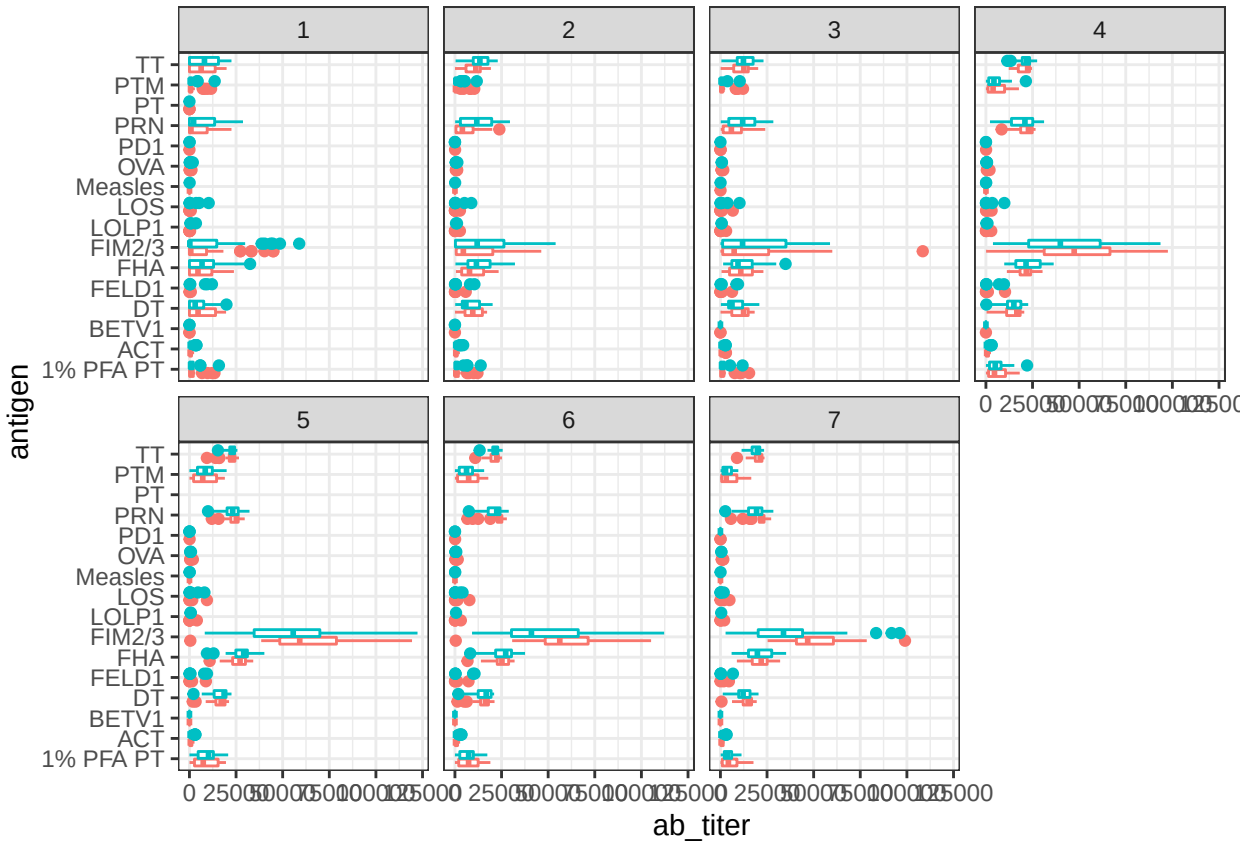
```
ggplot(ig1) +
  aes(ab_titer, antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow=2)
```



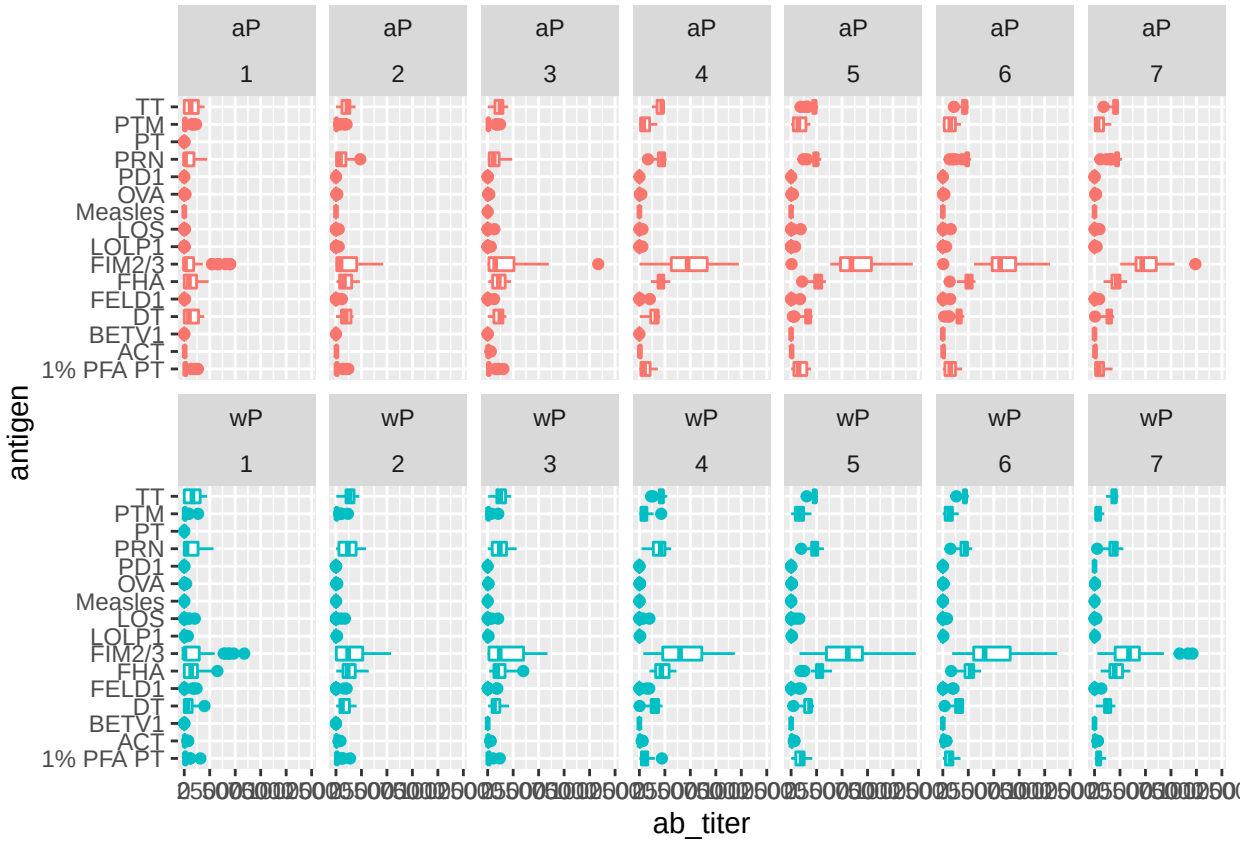
Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

FIM 2/3. I am not sure why, but possibly because these antigens are changing.

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

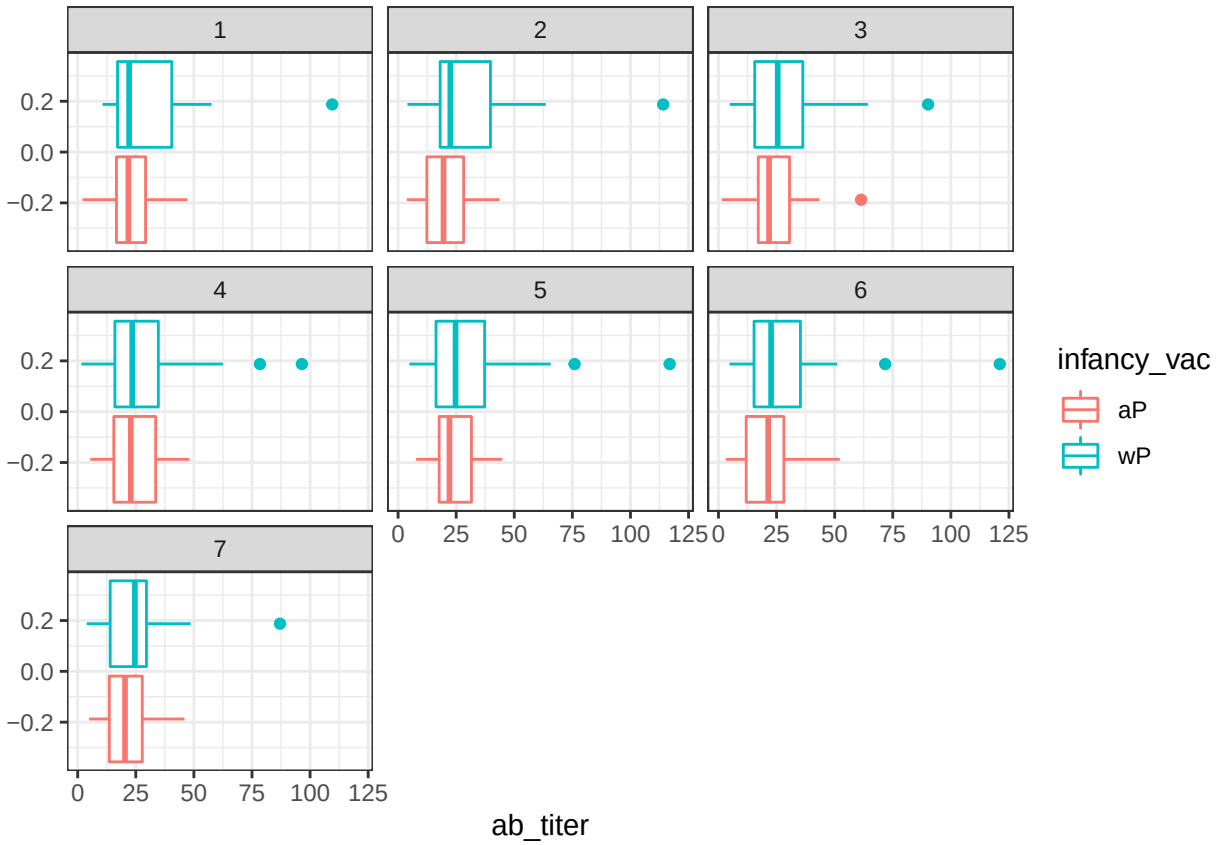



```
ggplot(ig1) +
  aes(ab_titer, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```

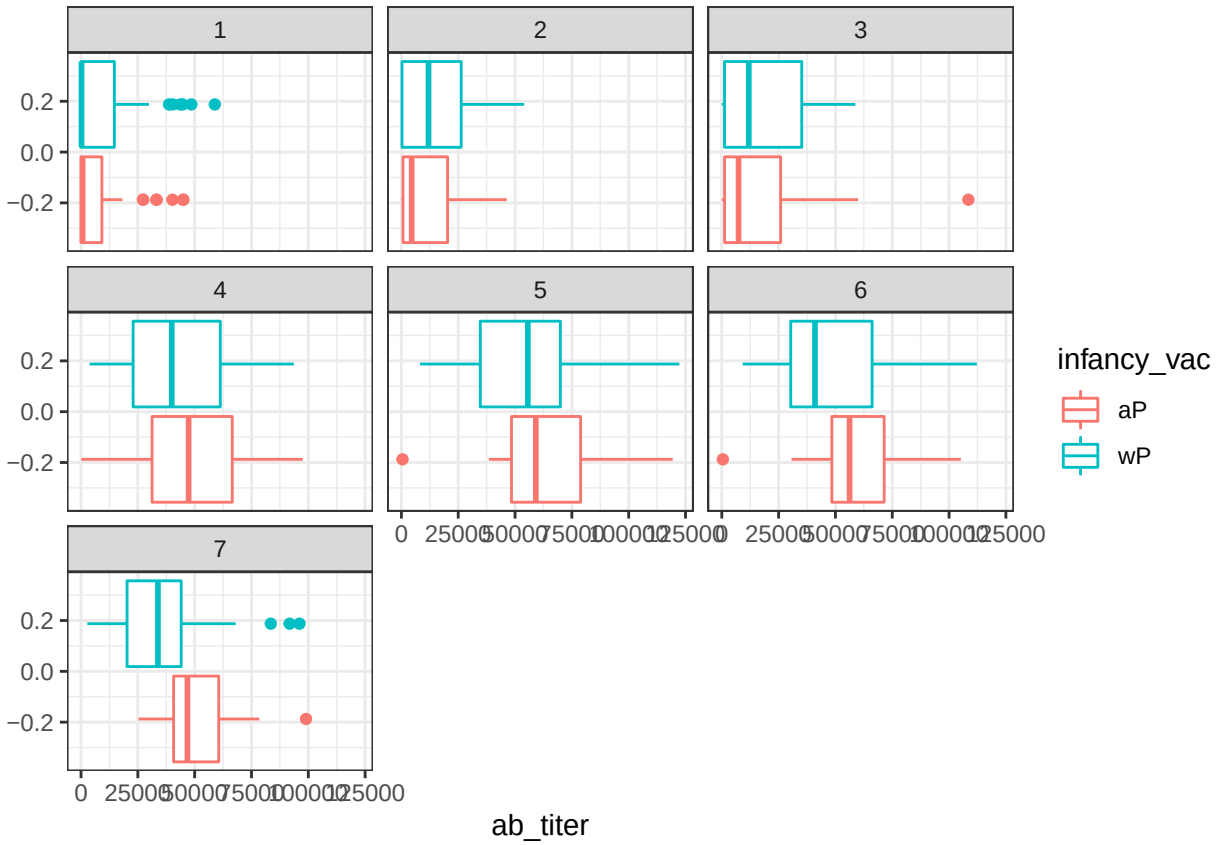


Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("Measles", that is not in our vaccines) and a clear antigen of interest ("FIM2/3", extra-cellular fimbriae proteins from B. pertussis that participate in substrate attachment).

```
filter(ig1, antigen=="Measles") %>%
  ggplot() +
  aes(ab_titer, col=infancy_vac) +
  geom_boxplot(show.legend = TRUE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```




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



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

FIM2/3 levels clearly rise over time and far exceed those of Measles. They also appear to peak at visit 5 and then decline. This trend appears similar for for wP and aP subjects.

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

No.

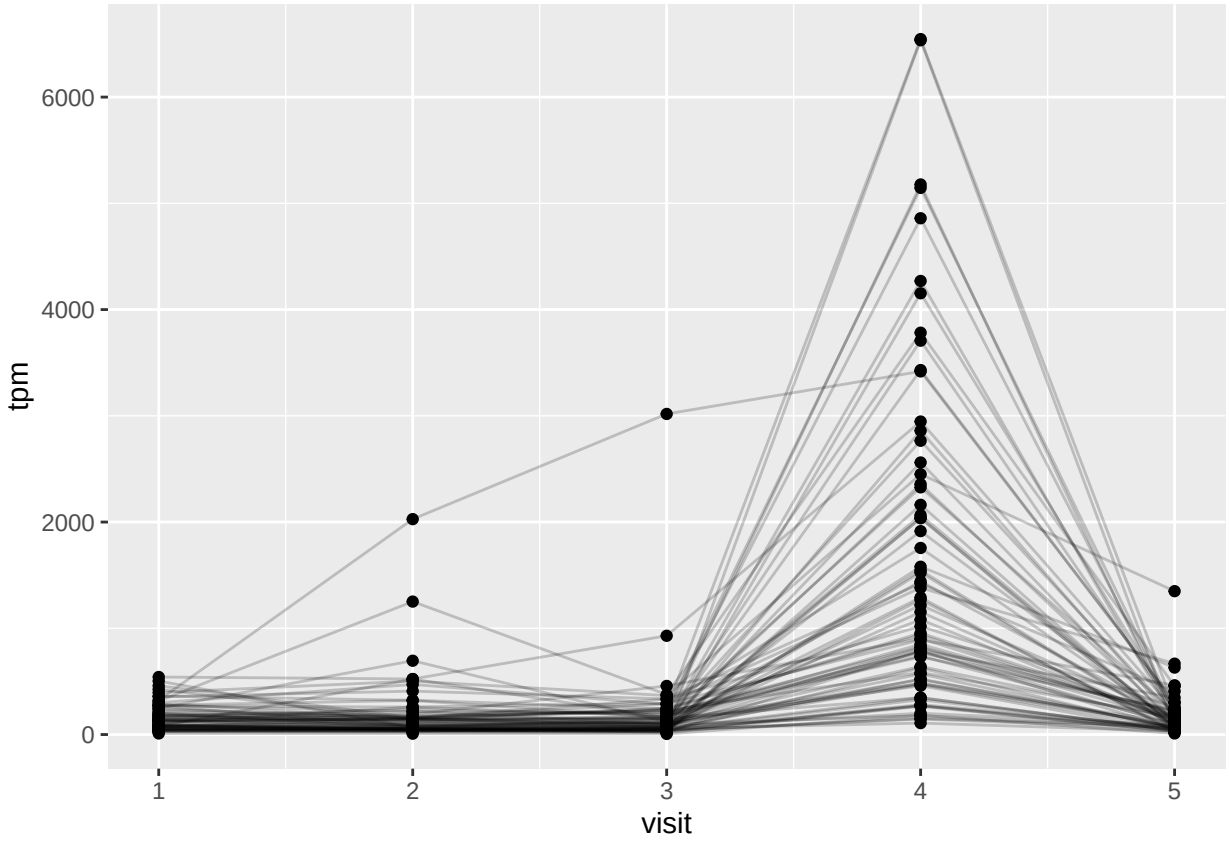
5. Obtaining CMI-PB RNASeq Data

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

rna <- read_json(url, simplifyVector = TRUE)

#meta <- inner_join(specimen, subject)
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)

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



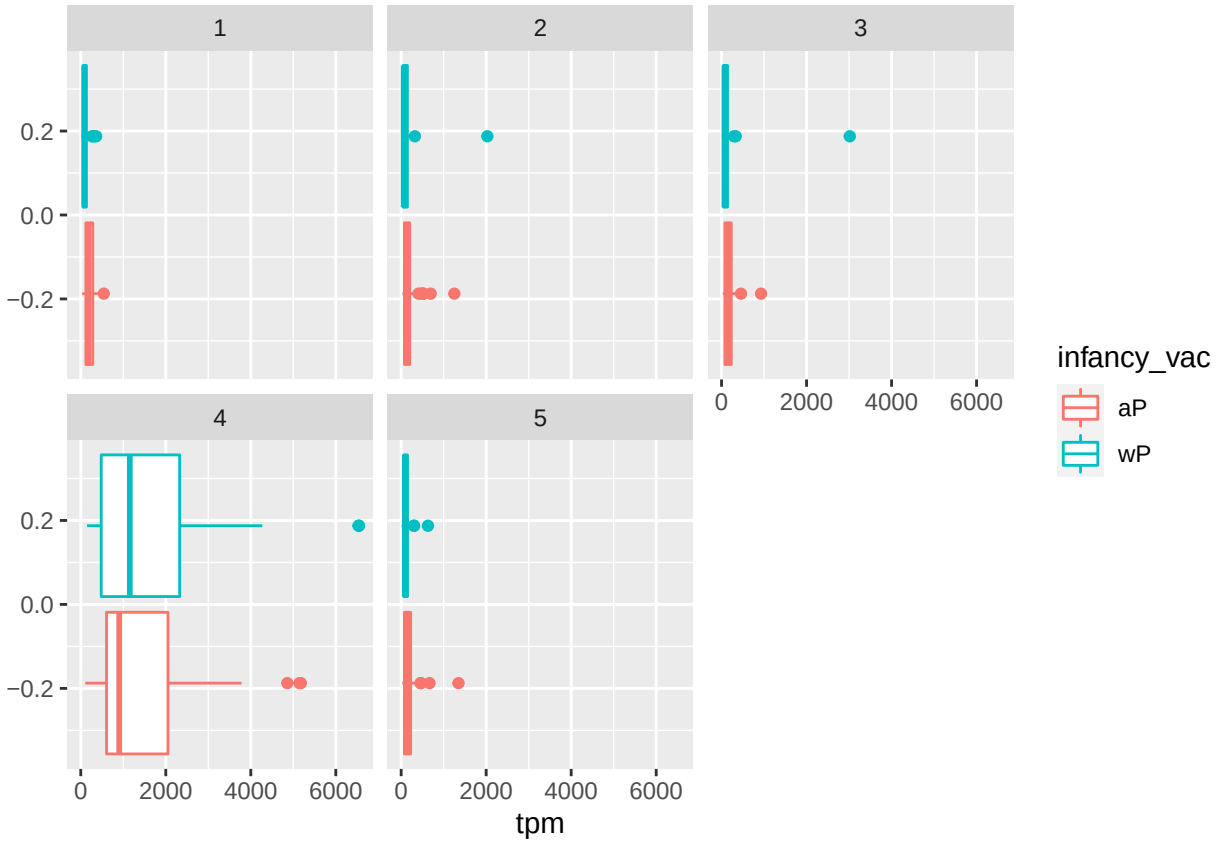
Q19.: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

It happens during visit 4

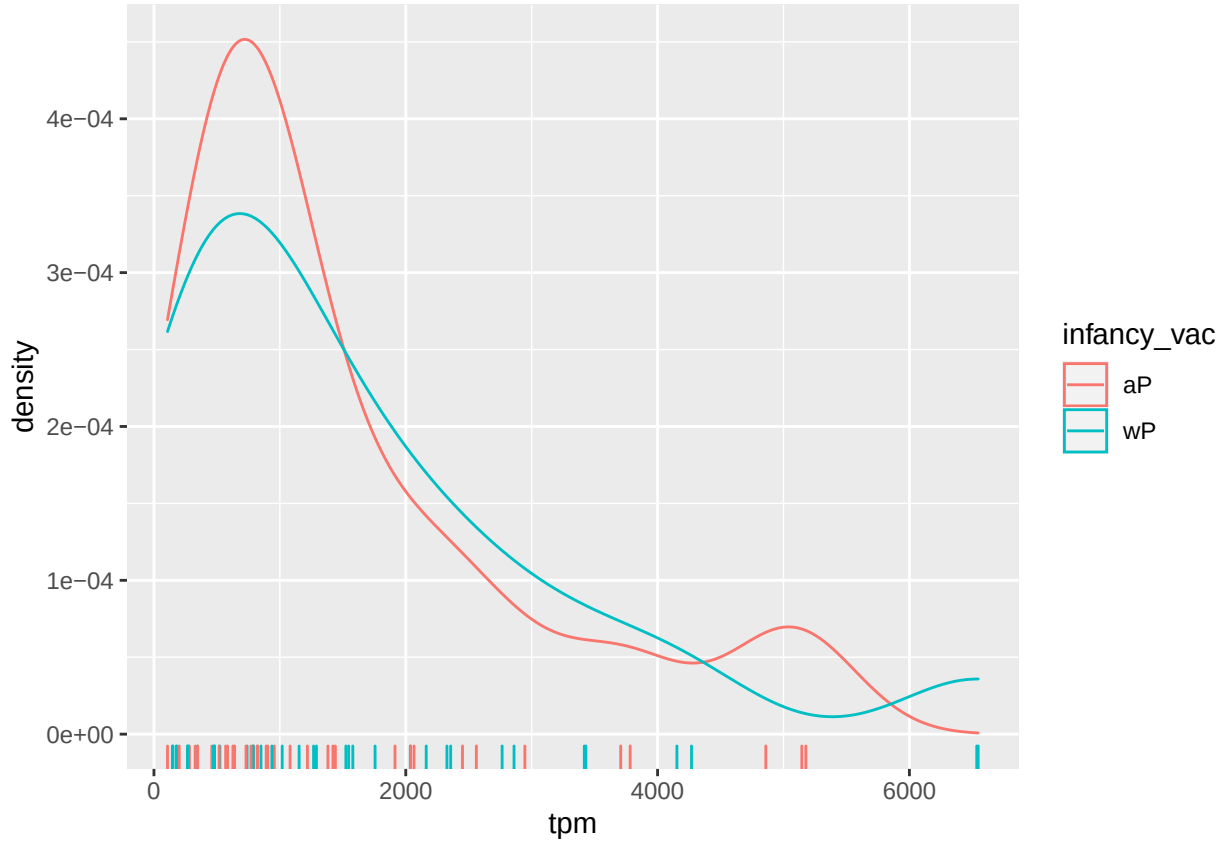
Q20. Does this pattern in time match the trend of antibody titer data? If not, why not?

No. For the antibody titer data, the antibody levels stay consistent over time rather than have a peak because antibodies are long lived.

```
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()
```




```
sessionInfo()
```

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19043)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] dplyr_1.0.8     rlang_1.0.2     lubridate_1.8.0 jsonlite_1.7.3
## [5] ggplot2_3.3.5
##
## loaded via a namespace (and not attached):
##  [1] highr_0.9         pillar_1.7.0      compiler_4.1.2    tools_4.1.2
##  [5] digest_0.6.27     evaluate_0.14     lifecycle_1.0.1   tibble_3.1.6
##  [9] gtable_0.3.0      pkgconfig_2.0.3   cli_3.1.1         DBI_1.1.2
```

```
## [13] rstudioapi_0.13   yaml_2.2.1         xfun_0.29          withr_2.4.3
## [17] stringr_1.4.0     knitr_1.37         generics_0.1.2     vctrs_0.3.8
## [21] grid_4.1.2        tidyselect_1.1.1   glue_1.6.1         R6_2.5.1
## [25] fansi_1.0.2       rmarkdown_2.11     farver_2.1.0       purrr_0.3.4
## [29] magrittr_2.0.2    scales_1.1.1       ellipsis_0.3.2     htmltools_0.5.1.1
## [33] assertthat_0.2.1  colorspace_2.0-2   labeling_0.4.2     utf8_1.2.2
## [37] stringi_1.7.6     munsell_0.5.0      crayon_1.5.0
```