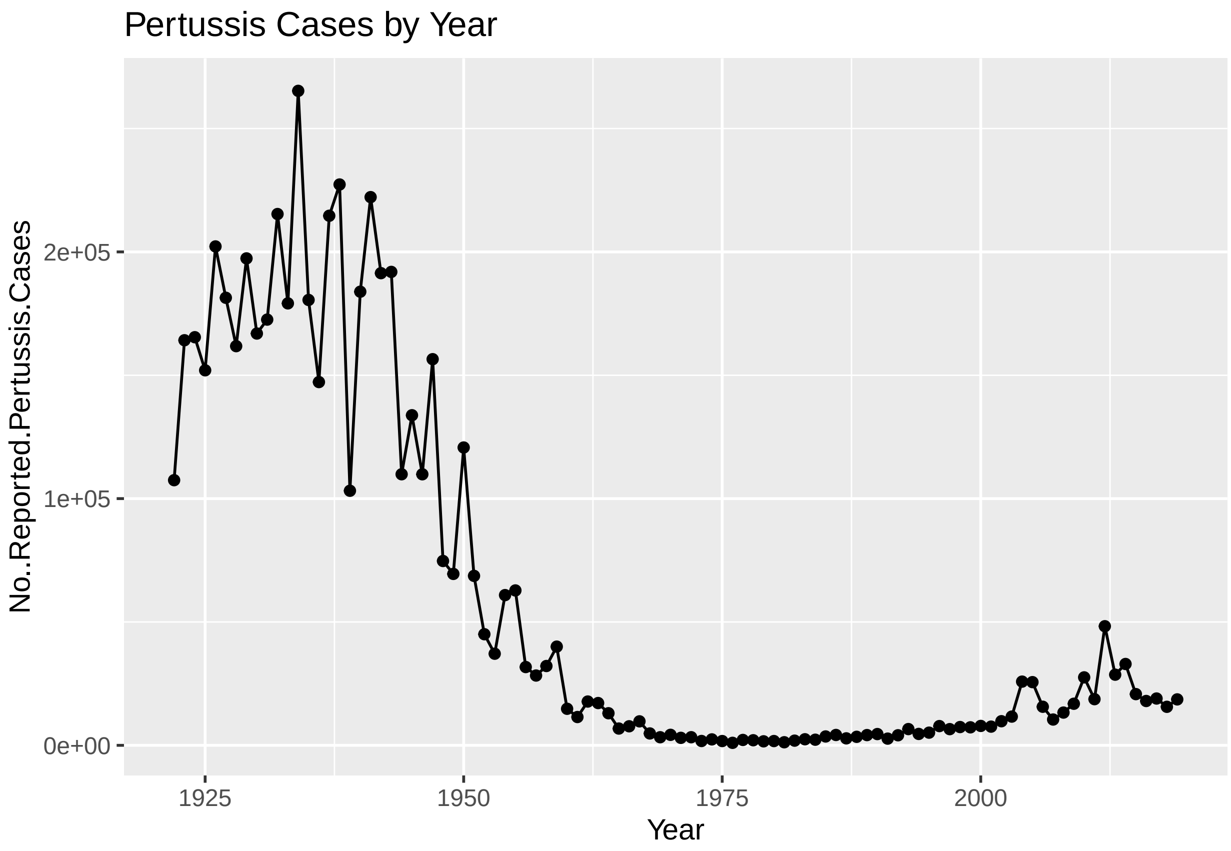
Class15

Ayse

2022-03-09

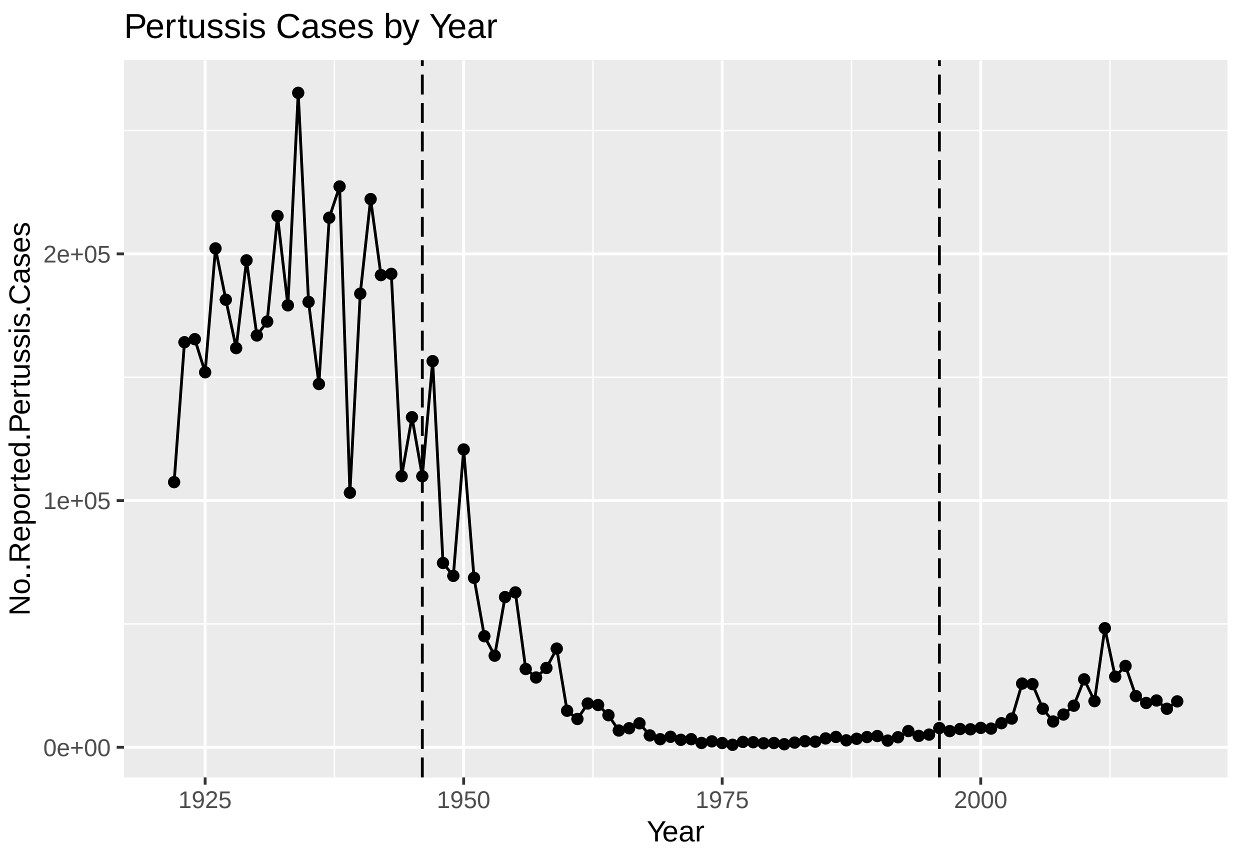
```
Q1. With the help of the R "addin" package datapasta assign the CDC pertussis case number data to a data
```

```
#install.packages("datapasta")
library(datapasta)
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)
ggplot(cdc) +
  aes(Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +
  labs(title="Pertussis Cases by Year")
```



Q2. Using the ggplot geom_vline() function add lines to your previous plot for the 1946 introduction of

```
ggplot(cdc) +
  aes(Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_vline(xintercept = 1946, linetype = "longdash") +
  geom_vline(xintercept = 1996, linetype = "longdash") +
  geom_line() +
  labs(title="Pertussis Cases by Year")
```



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

At the first introduction of the aP vaccine, there seemed to be a continuation of the minimal cases. However, slight increases in the number of cases can be seen after 2000. This may be due to a drop in vaccination rates, loss of immunity over the years for patients who were vaccinated before 2000, or more sensitive tests to detect pertussis being developed after the year 2000.

```
# 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 many aP and wP infancy vaccinated subjects are in the dataset?
```

47 for aP and 49 for wP.

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

```
##
## aP wP
## 47 49
```

Q5. How many Male and Female subjects/patients are in the dataset?
66 female and 30 male.

```
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...)?
The breakdown is as shown in the table below.

```
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
```

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? i. 35 ii. 25 iii. yes

```
library(lubridate)
```

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

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

```
#add column age showing the age of each subject
subject$age <- time_length(today() - ymd(subject$year_of_birth), "years")

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
```

```
#filter for individuals that were vaccinated with aP
ap <- subject %>% filter(infancy_vac == "aP")

mean_ap <- mean(ap$age)
mean_ap
```

```
## [1] 24.5026
```

```
#filter for individuals that were vaccinated with wP
wp <- subject %>% filter(infancy_vac == "wP")

mean_wp <- mean(wp$age)
mean_wp
```

```
## [1] 35.34705
```

Q8. Determine the age of all individuals at time of boost?
Ages are listed below. Average age of boost is around 25.

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

```
##  [1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481 35.84942 34.14921
##  [9] 20.56400 34.56263 30.65845 34.56263 19.56194 23.61944 27.61944 29.56331
## [17] 36.69815 19.65777 22.73511 32.26557 25.90007 23.90144 25.90007 28.91992
## [25] 42.92129 47.07461 47.07461 29.07324 21.07324 21.07324 28.15058 24.15058
## [33] 24.15058 21.14990 21.14990 31.20876 26.20671 32.20808 27.20876 26.20671
## [41] 21.20739 20.26557 22.26420 19.32375 21.32238 19.32375 19.32375 22.41752
## [49] 20.41889 21.41821 19.47707 23.47707 20.47639 21.47570 19.47707 35.65777
## [57] 33.65914 31.65777 25.73580 24.70089 28.70089 33.73580 19.73443 34.73511
## [65] 19.73443 28.73648 27.73443 19.81109 26.77344 33.81246 25.77413 19.81109
## [73] 18.85010 19.81109 31.81109 22.81177 31.84942 19.84942 18.85010 18.85010
## [81] 19.90691 18.85010 20.90897 19.04449 20.04381 19.90691 19.90691 19.00616
## [89] 19.00616 20.04381 20.04381 20.07940 21.08145 20.07940 20.07940 20.07940
```

```
mean(subject$boostage)
```

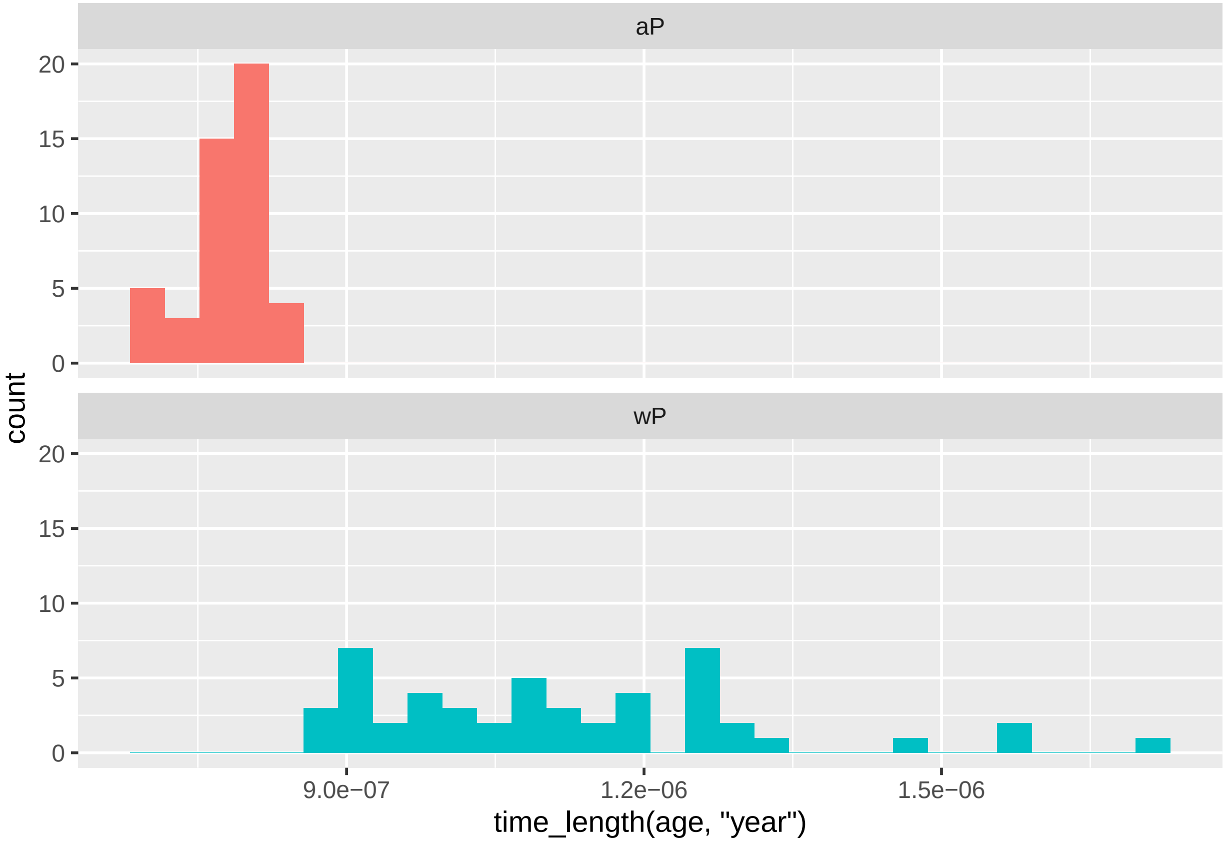
```
## [1] 25.60763
```

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

yes, there is no overlap

```r
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‘.
```



```r
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 al
```

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

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

```r
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 boostage
## 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
```

```
Q10. Now using the same procedure join meta with titer data so we can further analyze this data in terms
```

```
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?
```

Shown in the table below.

```
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?
```

there are much fewer visit 8 specimens compared to the other visits.

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

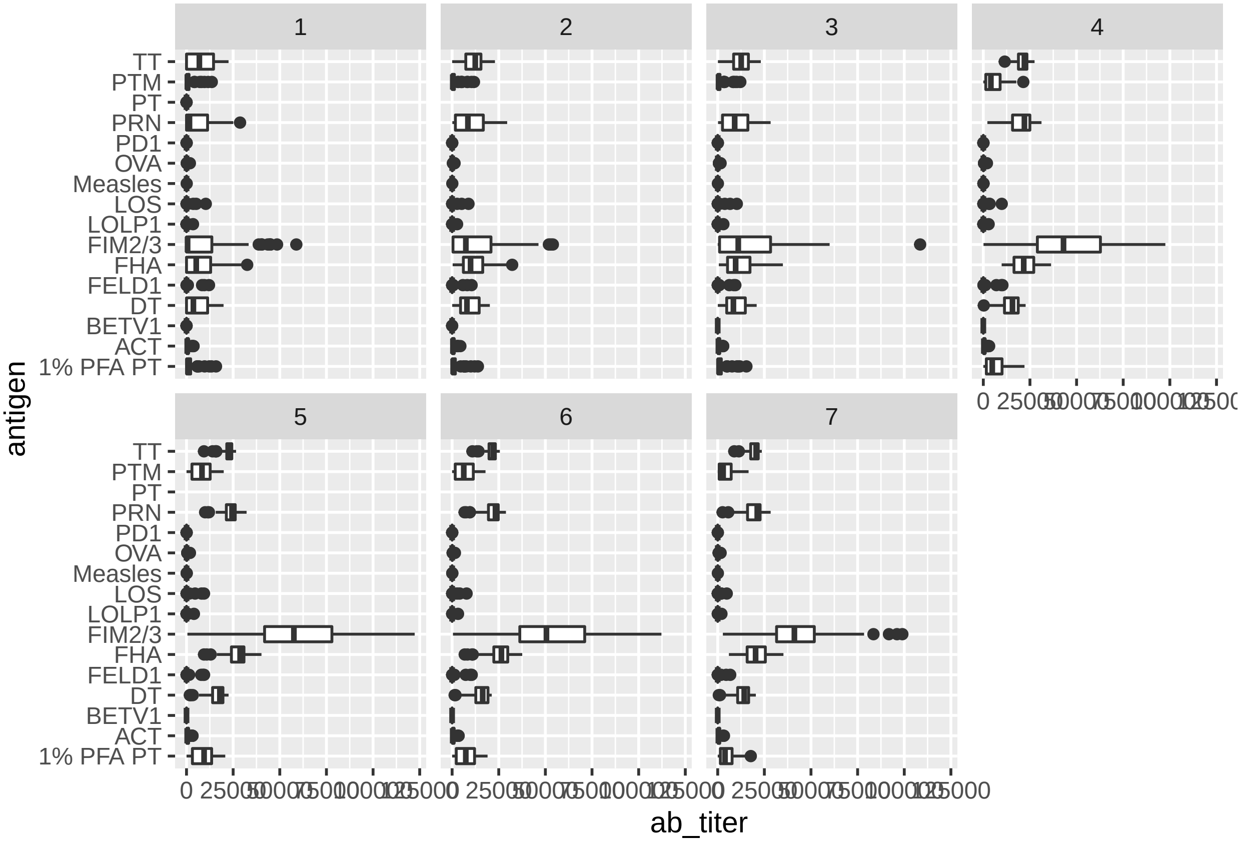
```
## 
##    1    2    3    4    5    6    7    8 
## 5795 4640 4640 4640 4640 4320 3920   80
```

```r
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 boostage
## 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
```

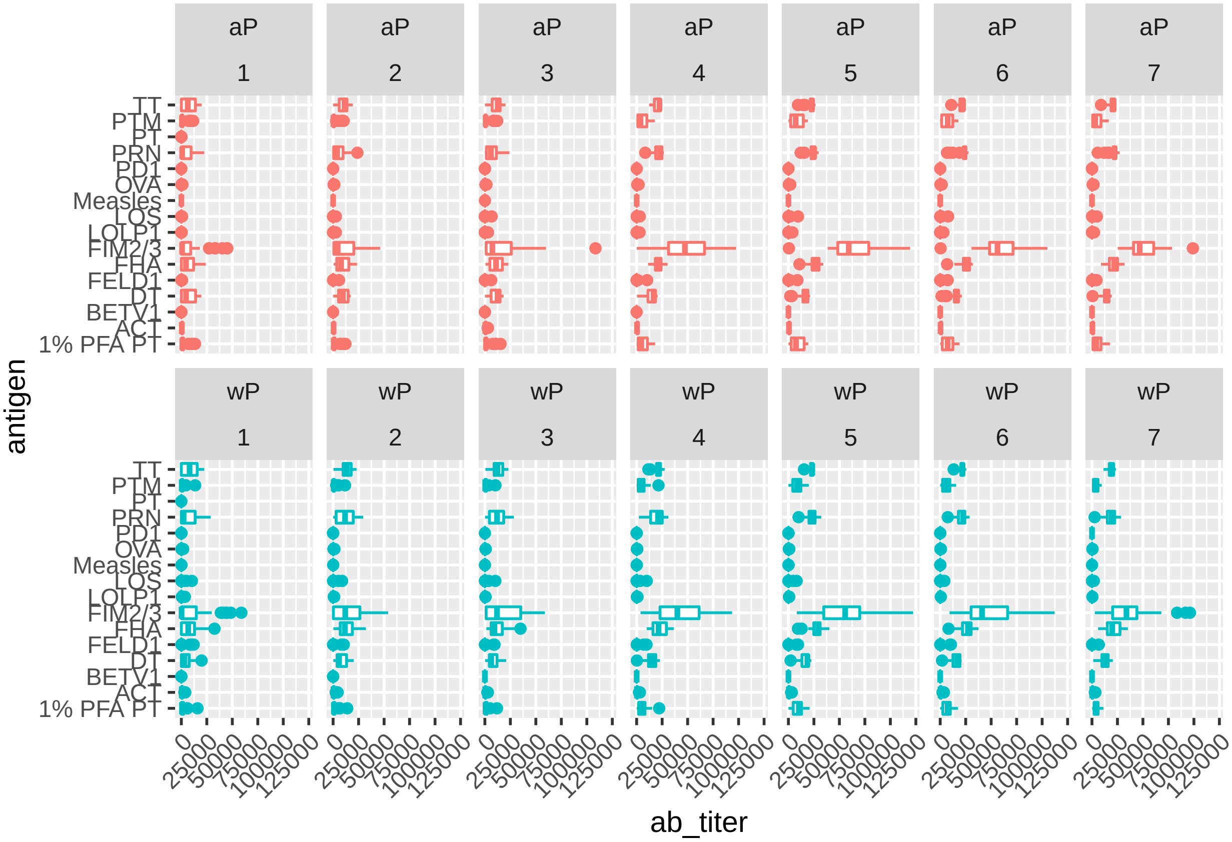
```
Q13. Complete the following code to make a summary boxplot of Ab titer levels for all antigens:
```

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



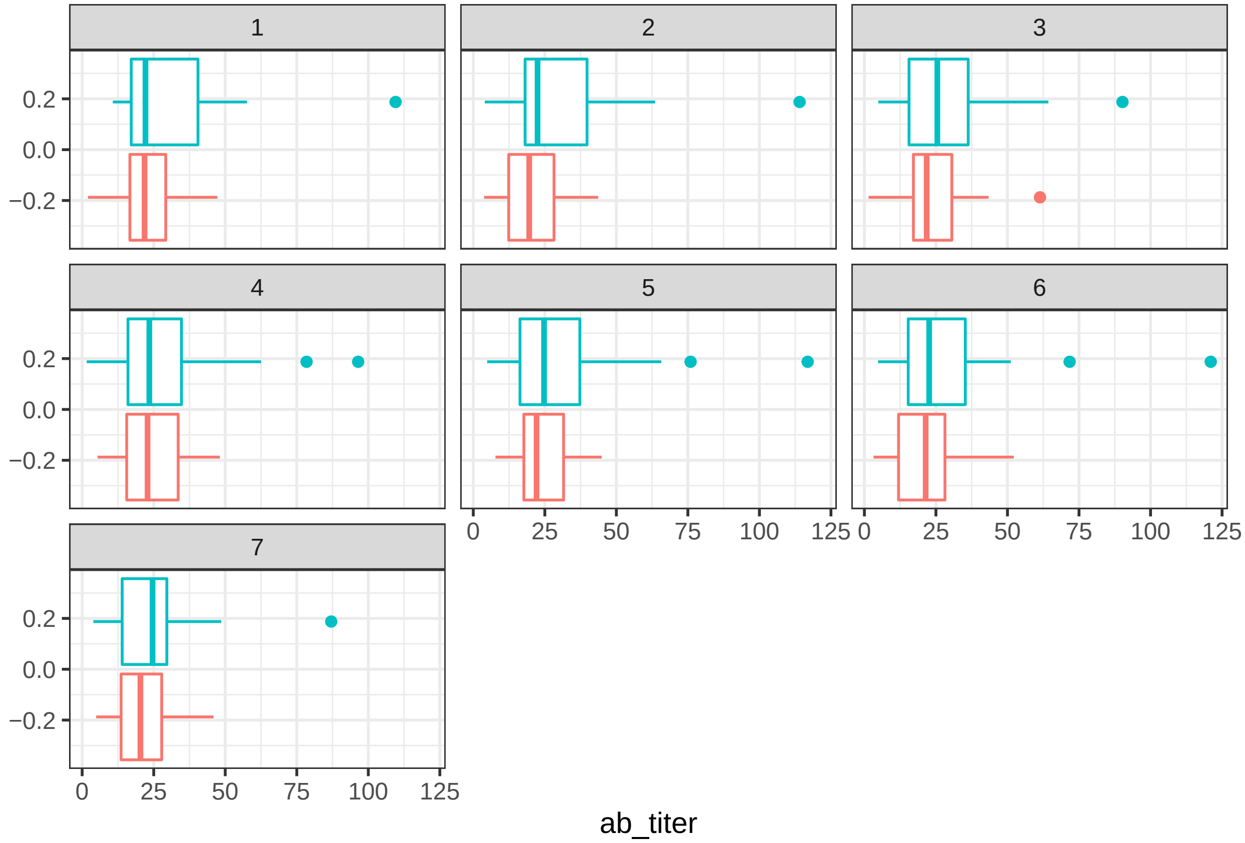
Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others? FIM2/3, FHA, PRN. These must have been low due to waning of immunity and upon boosting, increased.

```
ggplot(ig1) +
  aes(ab_titer, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(infancy_vac, visit), nrow=2) +
  theme(axis.text.x = element_text(angle = 45, hjust=1))
```

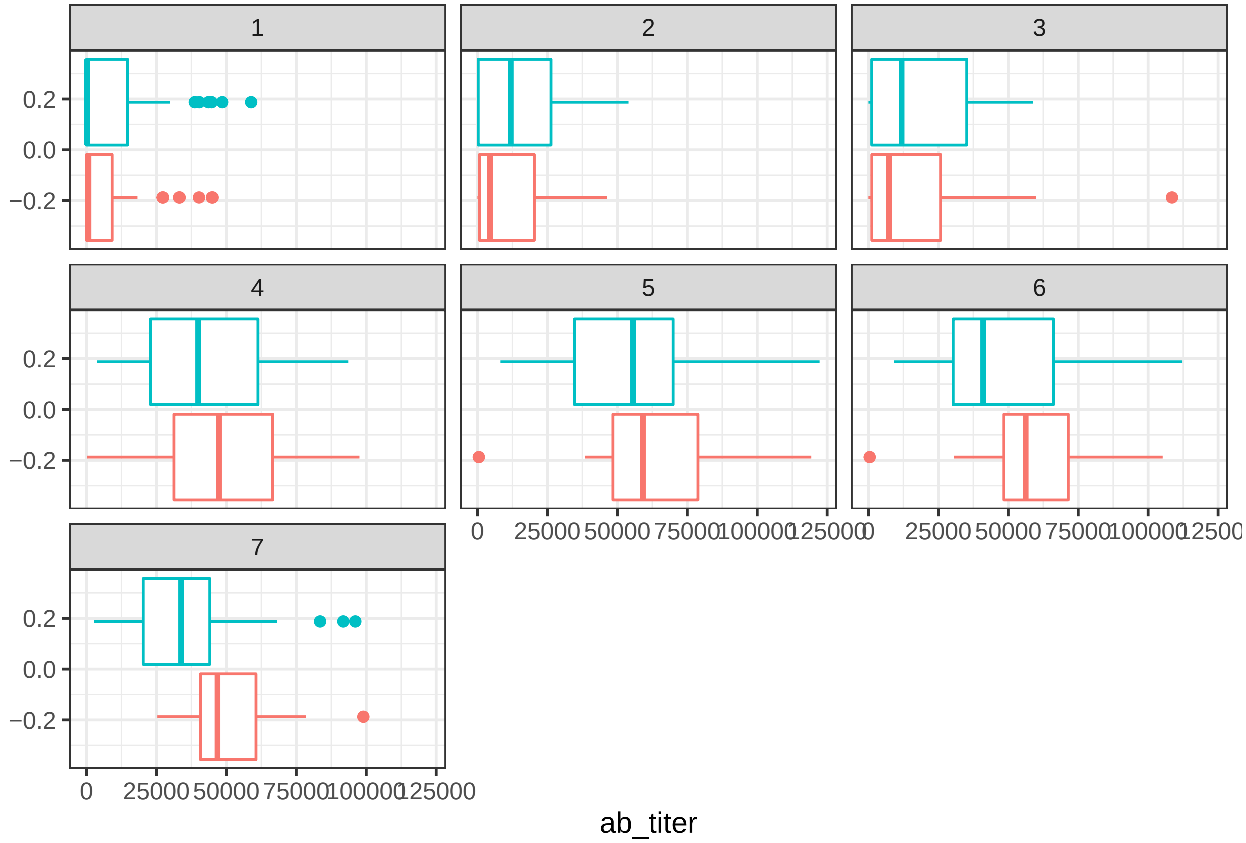


Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can

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



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



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

FIM2/3 titer levels are much higher than Measles and peaks at visit 5 for both wP and aP vaccinated subjects.

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

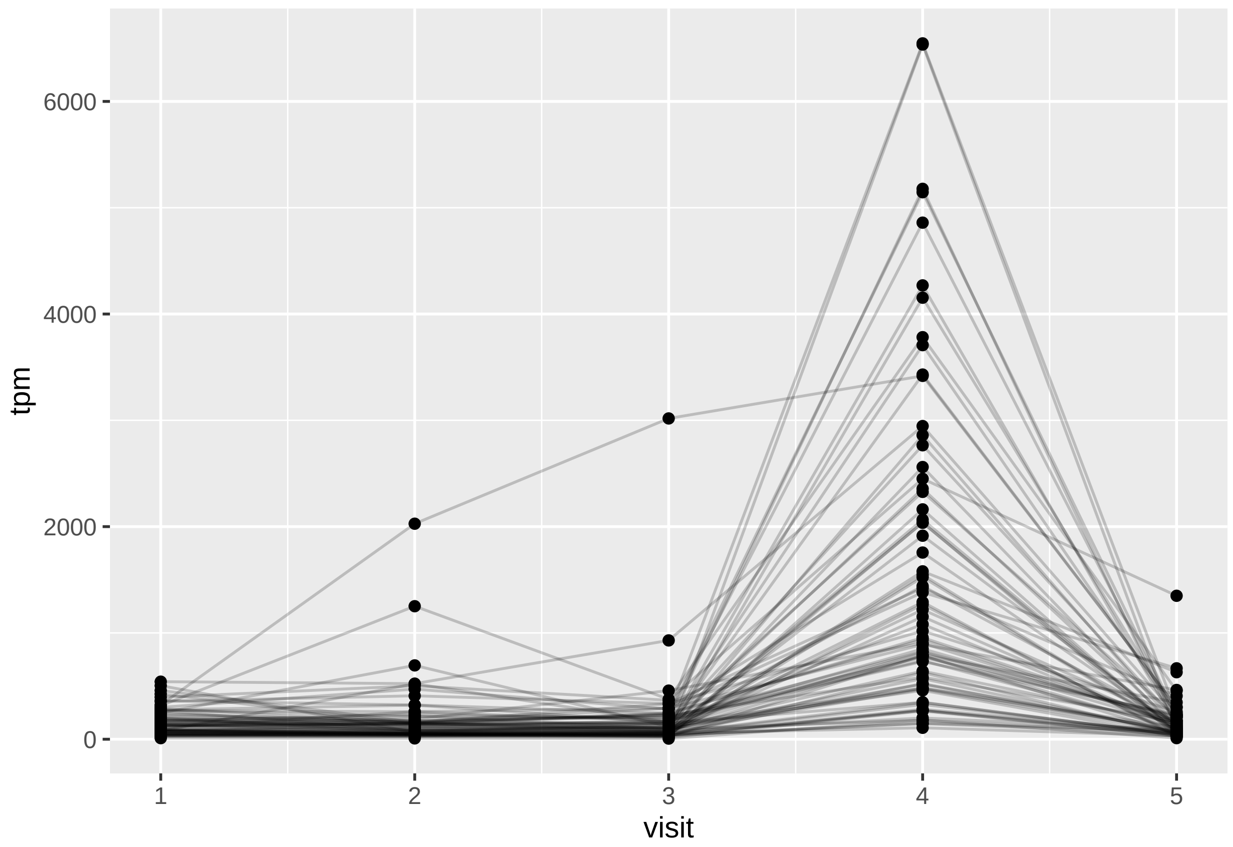
No.

```r
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).
```

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



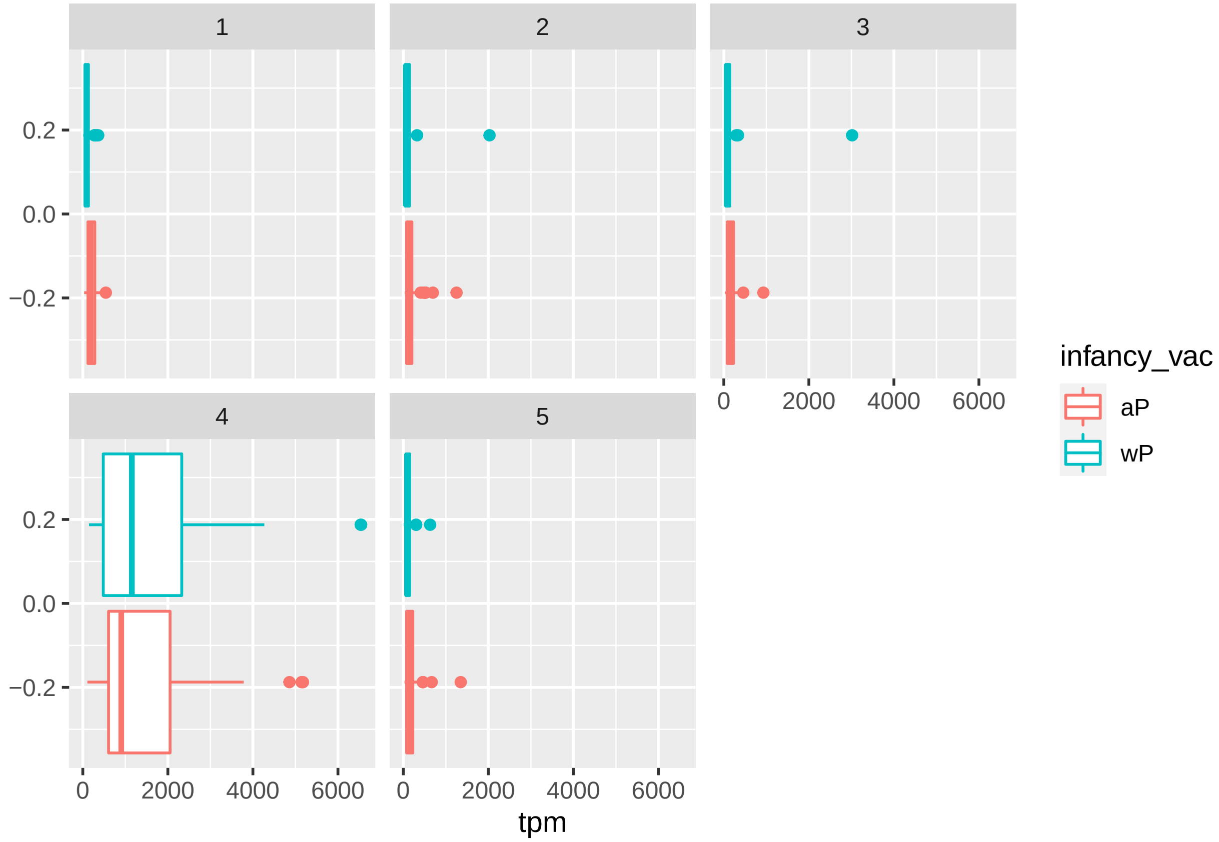
`Q19.: What do you notice about the expression of this gene (i.e. when is it at it’s maximum level)?`

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

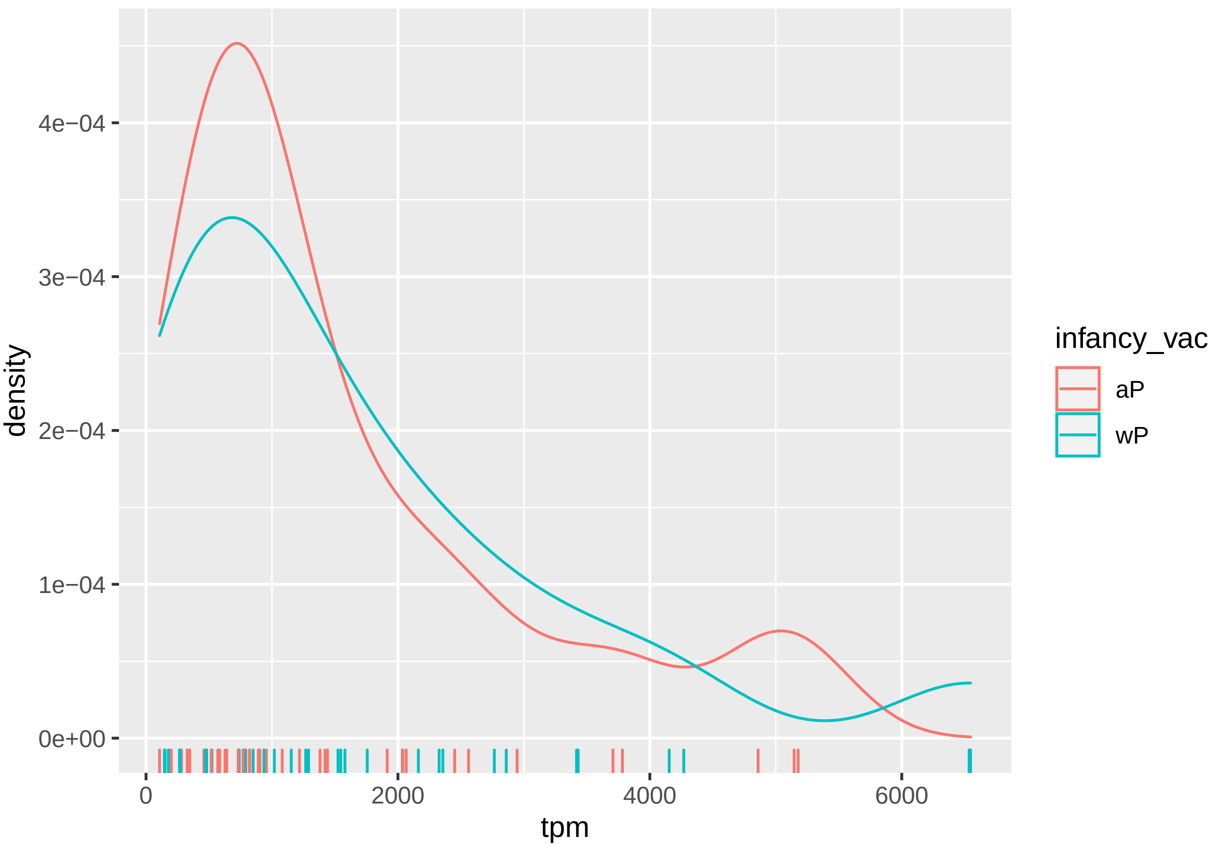
`Q20. Does this pattern in time match the trend of antibody titer data? If not, why not?`

No, because gene expression decreases after it peaks, whereas for the antibody titers, they stayed stable and did not decrease after they peaked.

```
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