Class_15_Miniproject

Samuel Do (PID:A15803613)

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 dat
a to a data frame called "cdc" and use ggplot to make a plot of cases numbers over time.
# 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)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

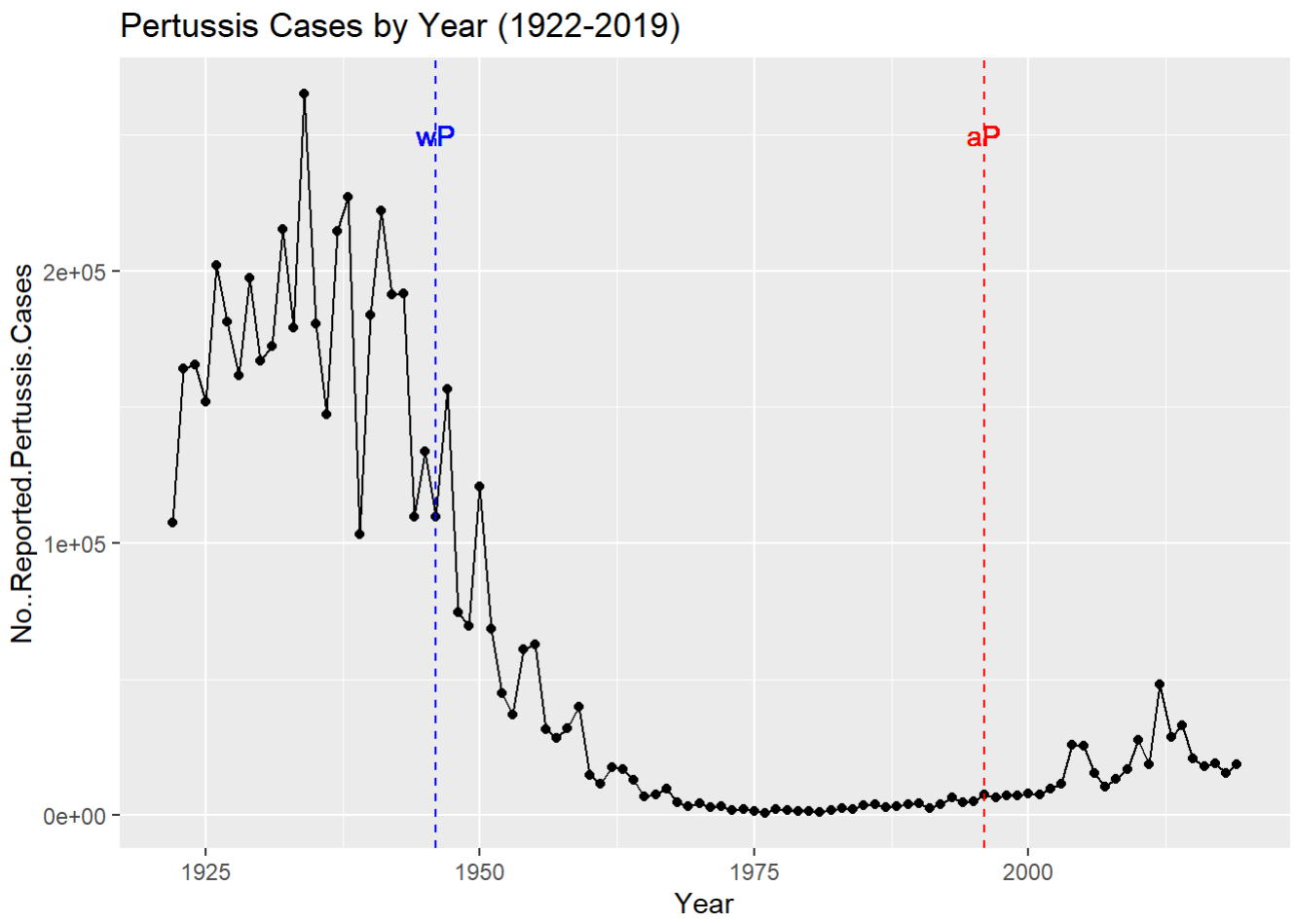
```
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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

2. A tale of two vaccines (wP & aP)

```
# [Q2] Using the ggplot geom_vline() function add lines to your previous plot for the 1946 intro
duction of the wP vaccine and the 1996 switch to aP vaccine (see example in the hint below). Wha
t do you notice?
ggplot(cdc, aes(x=Year, y=No..Reported.Pertussis.Cases)) +
  geom_point() +
  geom_line() +
  labs(title="Pertussis Cases by Year (1922-2019)") +
  geom_vline(xintercept = 1946, linetype = "dashed", color = "blue") +
  geom_text(aes(x=1946,label="wP", y=2.5e+05), color="blue") +
  geom_vline(xintercept = 1996, linetype = "dashed", color = "red") +
  geom_text(aes(x=1996,label="aP", y=2.5e+05), color="red")
```




```
# [Q3] Describe what happened after the introduction of the aP vaccine? Do you have a possible e
xplanation for the observed trend?
# This could be due to a newer variant of pertussis bacteria, which arises from bacterial evolut
ion.
```

3. Exploring CMI-PB data

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

```
##
## Attaching package: 'jsonlite'
```

```
## The following object is masked from 'package:purrr':
##
##     flatten
```

```
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
```

```
# There are 47 aP subjects and 49 wP subjects.
#[Q5] How many Male and Female subjects/patients are in the dataset?
table(subject$biological_sex)
```

```
##
## Female   Male
##     66     30
```

```
# There are 66 female subjects and 30 male subjects.
#[Q6] What is the breakdown of race and biological sex (e.g. number of Asian females, White male
s 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
```

```
library(lubridate)
```

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

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

```
library(dplyr)
# [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?
today()
```

```
## [1] "2022-03-09"
```

```
wP_subject <- subject %>% filter(infancy_vac=="wP")
aP_subject <- subject %>% filter(infancy_vac=="aP")
wP.age <- time_length(today()-ymd(wP_subject$year_of_birth), "years")
aP.age <- time_length(today()-ymd(aP_subject$year_of_birth), "years")
summary(wP.age)
```

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

```
summary(aP.age)
```

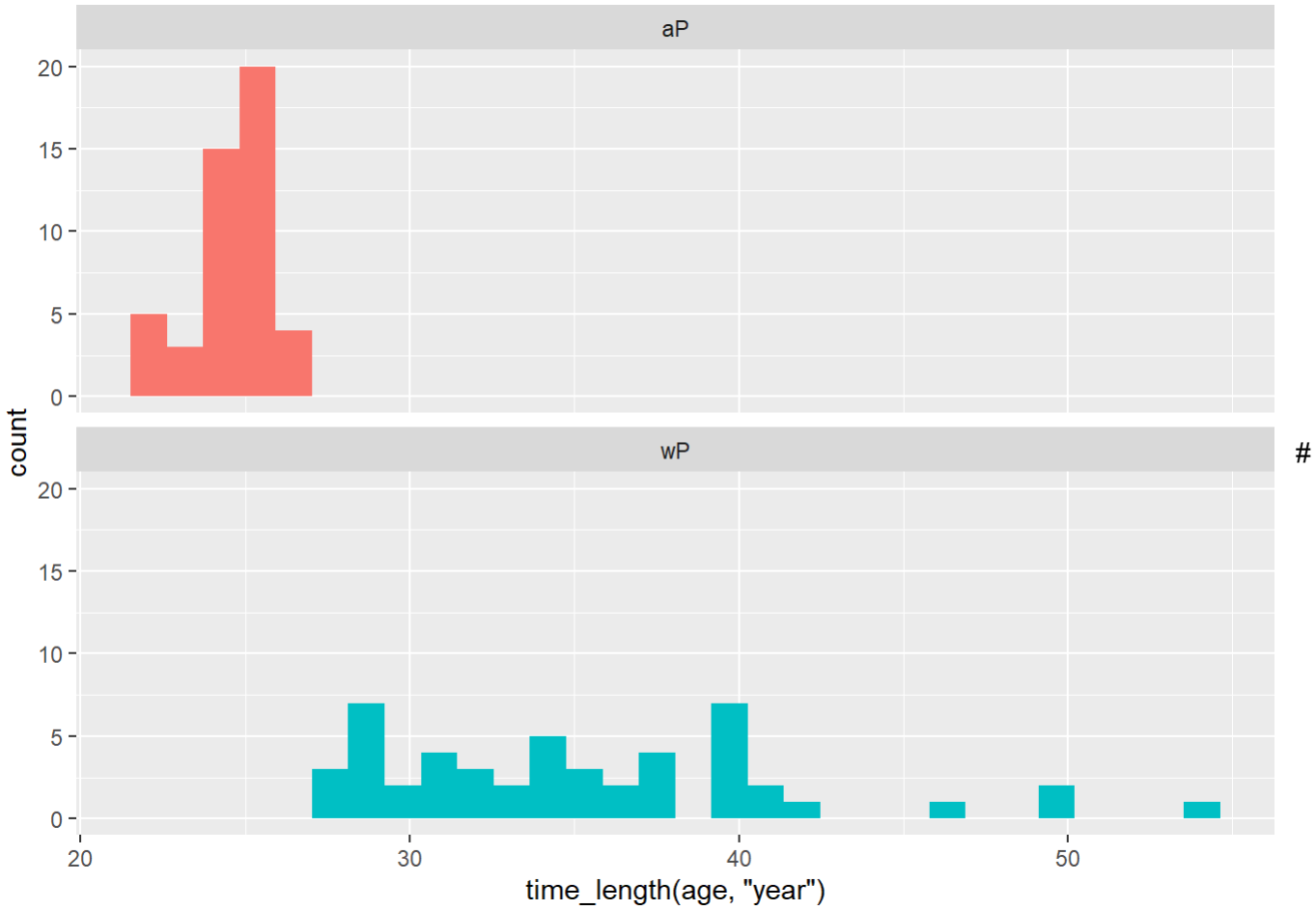
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.18   24.18   25.18   24.50   25.18   26.18
```

```
# The average age of the wP subjects was 35.34 years, while the average for aP subjects was 24.50 years. The ages of the two groups of subjects is significantly different as the IQR of the wP subjects' age does not overlap with that of the aP subjects' age.
# [Q8] Determine the age of all individuals at time of boost?
time_length(ymd(subject$date_of_boost)-ymd(subject$year_of_birth), "years")
```

```
##  [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
```

```
# [Q9] With the help of a faceted boxplot (see below), do you think these two groups are signifi
cantly different?
age <- today()-ymd(subject$year_of_birth)
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`.
```



#

Joining multiple tables

```
# 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)
head(specimen)
```

```
##   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
## 1                             0         Blood     1
## 2                           736         Blood    10
## 3                             1         Blood     2
## 4                             3         Blood     3
## 5                             7         Blood     4
## 6                            14         Blood     5
```

```
# [Q9] Complete the code to join specimen and subject tables to make a new merged data frame con
taining all specimen records along with their associated subject details:
meta <- inner_join(specimen, subject)
```

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

```
dim(meta)
```

```
## [1] 729  13
```

```
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
```

```
# [Q10] Now using the same procedure join meta with titer data so we can further analyze this da
ta in terms of time of visit aP/wP, male/female etc.
abdata <- inner_join(titer, meta)
```

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

```
dim(abdata)
```

```
## [1] 32675    19
```

```
head(abdata)
```

```
##   specimen_id isotype is_antigen_specific antigen   ab_titer  unit
## 1           1     IgE               FALSE   Total 1110.21154 UG/ML
## 2           1     IgE               FALSE   Total 2708.91616 IU/ML
## 3           1     IgG                TRUE      PT   68.56614 IU/ML
## 4           1     IgG                TRUE     PRN  332.12718 IU/ML
## 5           1     IgG                TRUE     FHA 1887.12263 IU/ML
## 6           1     IgE                TRUE     ACT    0.10000 IU/ML
##   lower_limit_of_detection subject_id actual_day_relative_to_boost
## 1                      NaN          1                           -3
## 2                29.170000          1                           -3
## 3                 0.530000          1                           -3
## 4                 1.070000          1                           -3
## 5                 0.064000          1                           -3
## 6                 2.816431          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
```

```
# [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
```

```
# There are 6698 IgE isotypes, 1413 IgG isotypes, 6141 IgG1 isotypes, 6141 IgG2 isotypes, 6141 I
gG3 isotypes, and 6141 IgG4 isotypes.
# [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
```

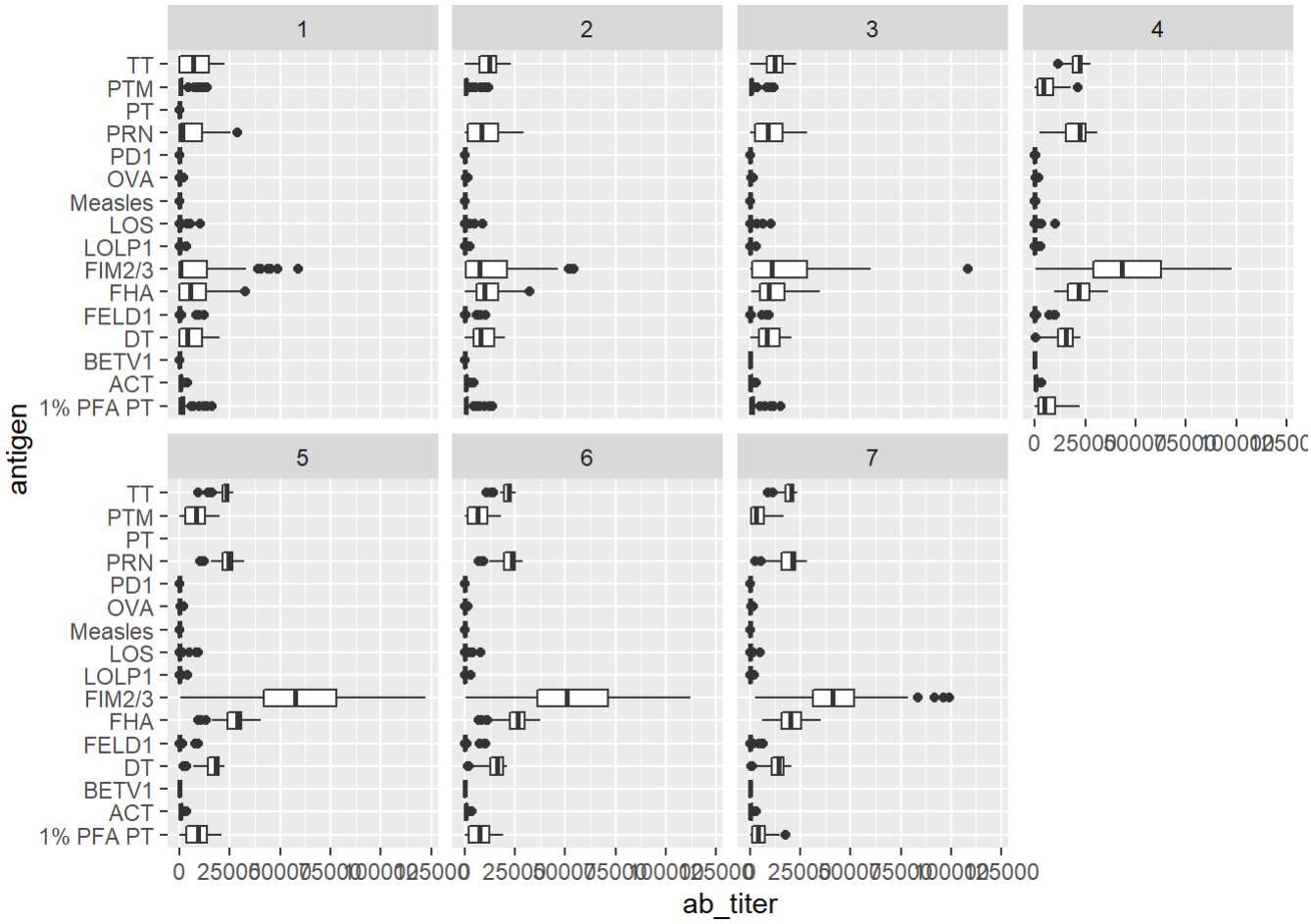
```
# There is a very small number of visit 8 specimens compared to the other visits.
```

4. Examine IgG1 Ab titer levels

```
# Filter IgG1 isotypes and exlude visit 8 subjects
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
```

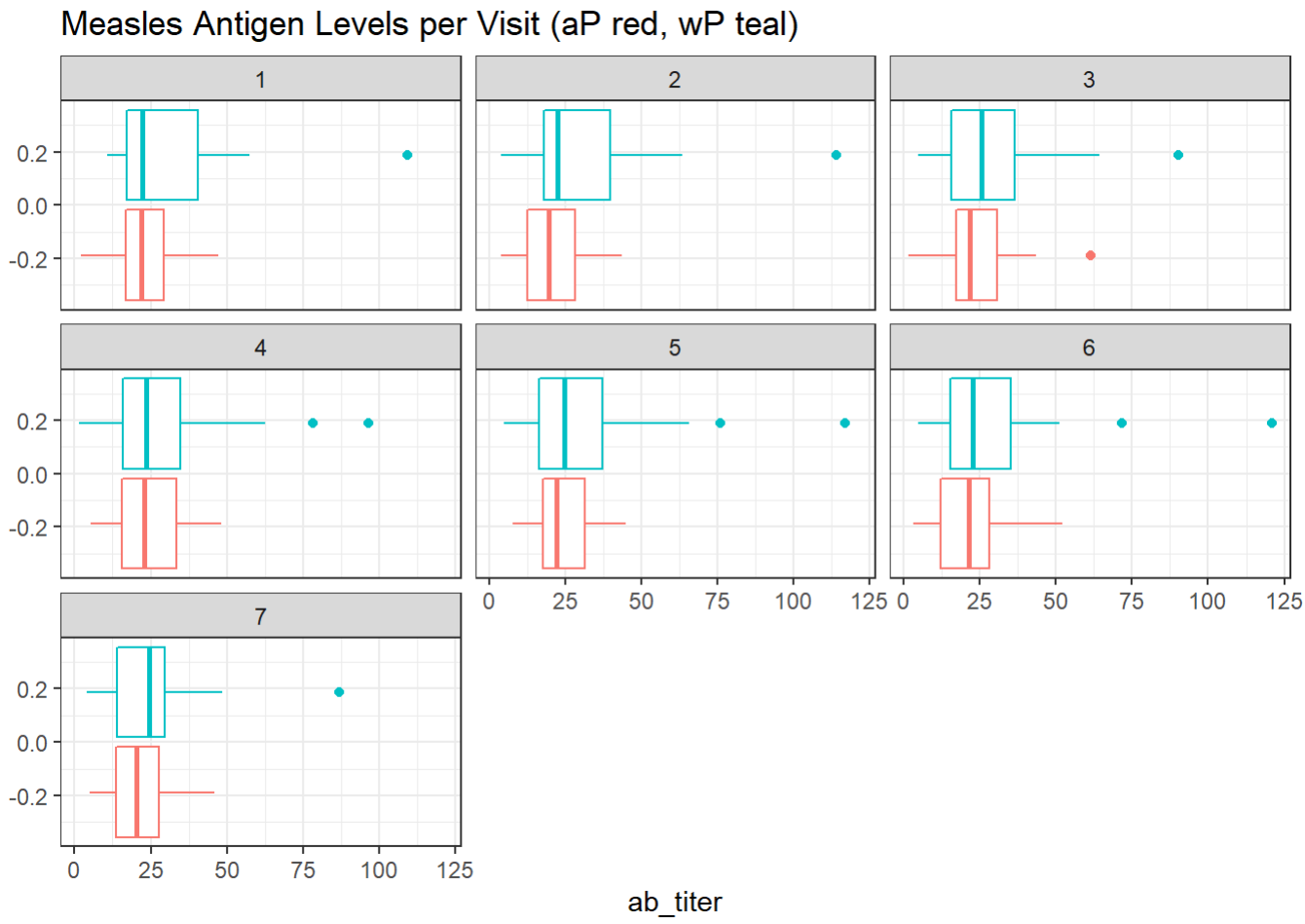
```
#[Q13] Complete the following code to make a summary boxplot of Ab titer levels for all antigen
s:
ggplot(ig1) +
  aes(ab_titer, antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow=2)
```



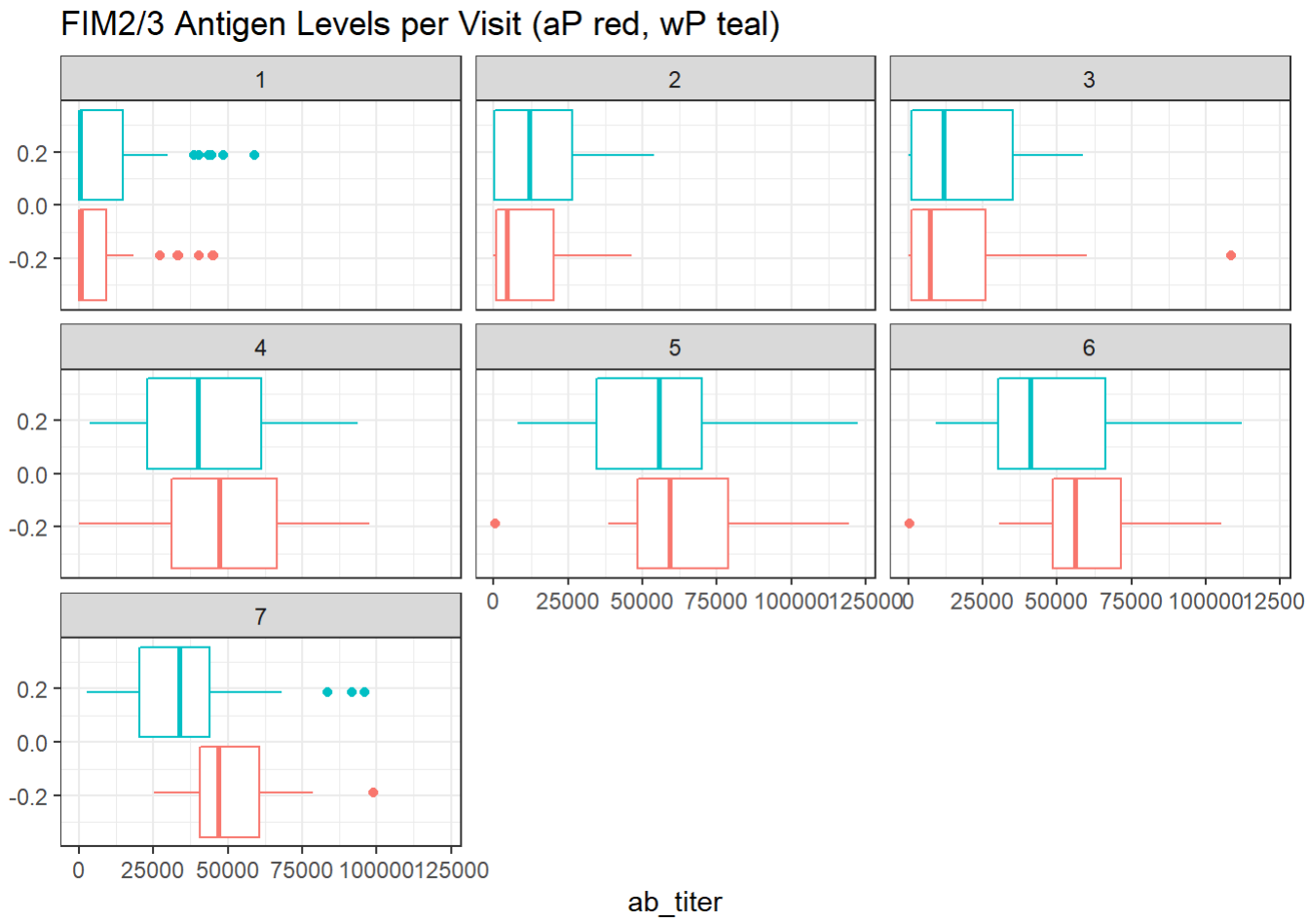
#[Q14] What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?
 # Antigen FIM2/3 is likely the antigen that is produced by bacteria that causes pertussis. Therefore, with the introduction of the vaccines, IgG1 antibodies are more likely to recognize these antigens, resulting in more immunity.

#[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 = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw() +
  labs(title="Measles Antigen Levels per Visit (aP red, wP teal)")
```




```
filter(ig1, antigen=="FIM2/3") %>%
  ggplot() +
  aes(ab_titer, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw() +
  labs(title="FIM2/3 Antigen Levels per Visit (aP red, wP teal)")
```




```
#[Q16] What do you notice about these two antigens time course and the FIM2/3 data in particula
r?
# The FIM2/3 antigen levels increase by a significant amount compared to the measles antigen lev
els, which stayed relative the same throughout all visits. The FIM2/3 antigen levels peak at vis
it 5 and decline by visit 6 onwards.
#[Q17] Do you see any clear difference in aP vs. wP responses?
# While there is no statistically significant difference between aP and wP responses due to the
 overlap in IQRs in all visits, FIM2/3 antigen levels were relatively greater with the aP vaccin
e after visit 3.
```

5. Obtaining CMI-PB RNASeq data

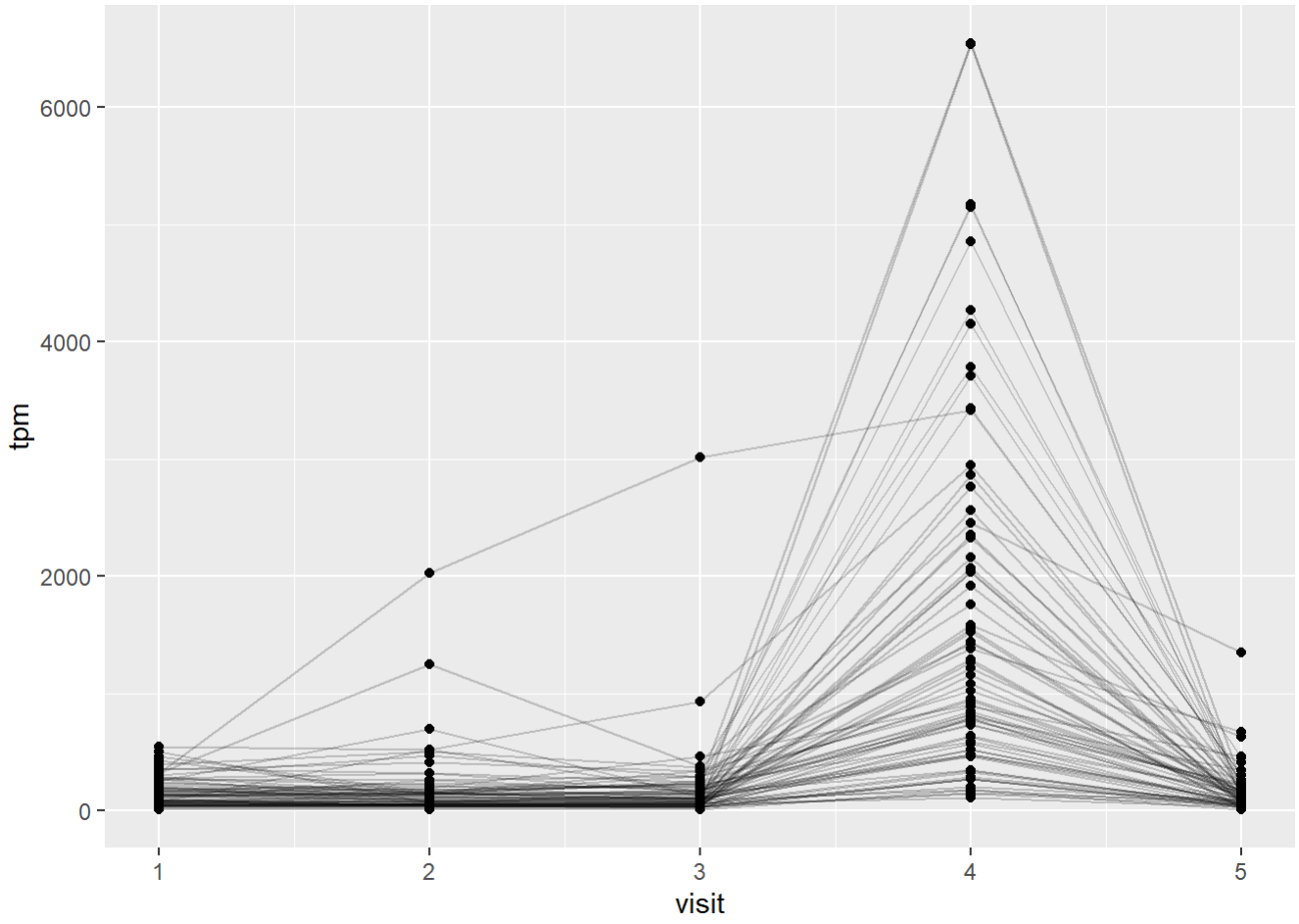
```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.7"
rna <- read_json(url, simplifyVector = TRUE)
#Join rna data with meta data
ssrna <- inner_join(rna, meta)
```

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

```
head(ssrna)
```

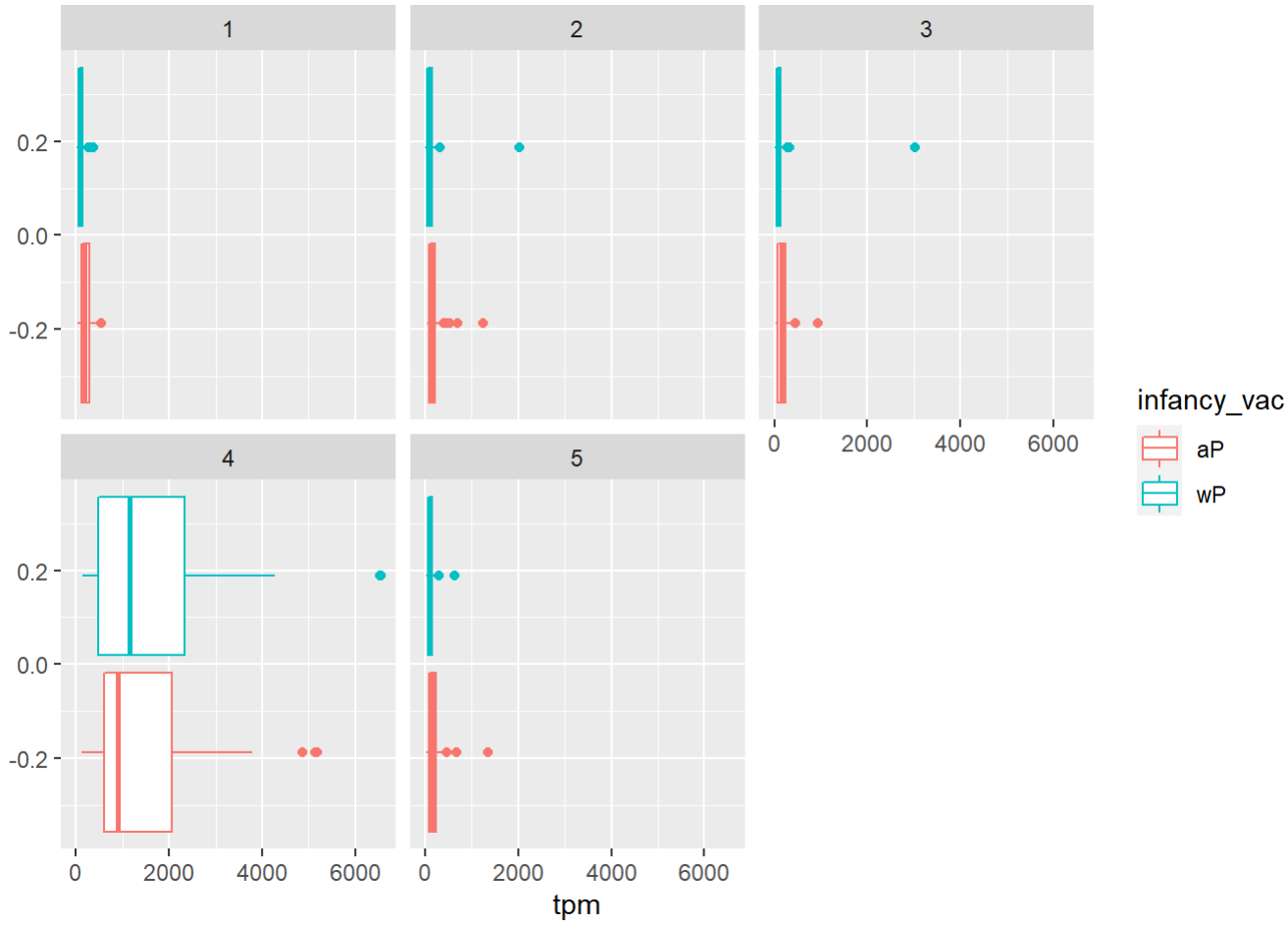
```
##   versioned_ensembl_gene_id specimen_id raw_count      tpm subject_id
## 1         ENSG00000211896.7         344     18613  929.640         44
## 2         ENSG00000211896.7         243      2011  112.584         31
## 3         ENSG00000211896.7         261      2161  124.759         33
## 4         ENSG00000211896.7         282      2428  138.292         36
## 5         ENSG00000211896.7         345     51963 2946.136         44
## 6         ENSG00000211896.7         244     49652 2356.749         31
##   actual_day_relative_to_boost planned_day_relative_to_boost specimen_type
## 1                            3                             3         Blood
## 2                            3                             3         Blood
## 3                           15                            14         Blood
## 4                            1                             1         Blood
## 5                            7                             7         Blood
## 6                            7                             7         Blood
##   visit infancy_vac biological_sex              ethnicity               race
## 1     3          aP         Female     Hispanic or Latino More Than One Race
## 2     3          wP         Female Not Hispanic or Latino              Asian
## 3     5          wP           Male     Hispanic or Latino More Than One Race
## 4     2          aP         Female     Hispanic or Latino              White
## 5     4          aP         Female     Hispanic or Latino More Than One Race
## 6     4          wP         Female Not Hispanic or Latino              Asian
##   year_of_birth date_of_boost   study_name
## 1    1998-01-01    2016-11-07 2020_dataset
## 2    1989-01-01    2016-09-26 2020_dataset
## 3    1990-01-01    2016-10-10 2020_dataset
## 4    1997-01-01    2016-10-24 2020_dataset
## 5    1998-01-01    2016-11-07 2020_dataset
## 6    1989-01-01    2016-09-26 2020_dataset
```

```
#[Q18] Make a plot of the time course of gene expression for IGHG1 gene (i.e. a plot of visit v
s. tpm).
ggplot(ssrna) +
  aes(visit, tpm, group=subject_id) +
  geom_point() +
  geom_line(alpha=0.2)
```

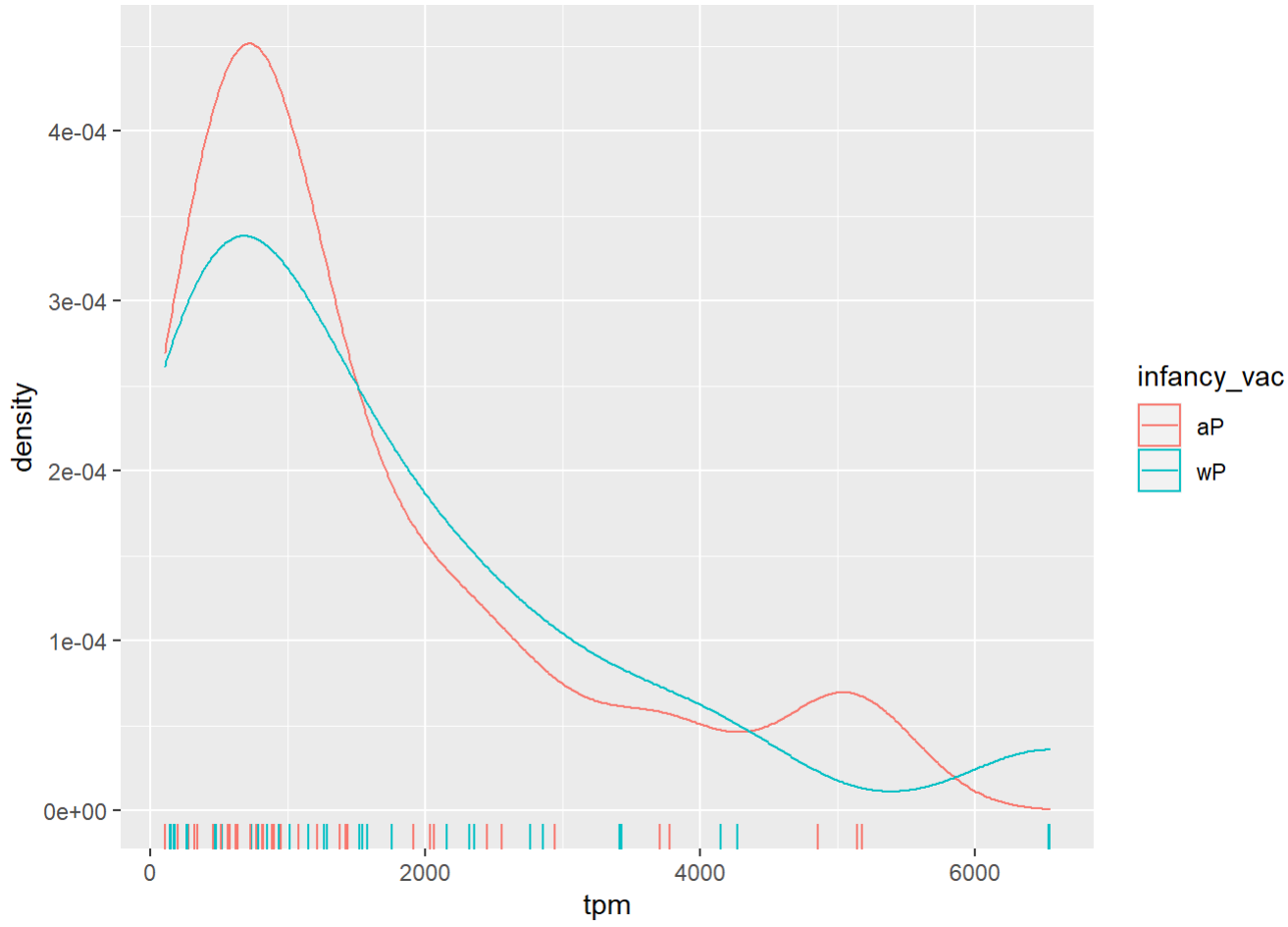


#[Q19] What do you notice about the expression of this gene (i.e. when is it at its maximum level)?
 # Expression of this gene is relatively low and suddenly peaks by visit 4, but then drops by visit 5.
#[Q20] Does this pattern in time match the trend of antibody titer data? If not, why not?
 # This pattern does not match the trend of antibody titer data as antibodies as the general expression of the gene would be more consistent and not suddenly peak and drop. This is because antibodies are long-lasting, which is not shown in this plot.

```
# Digging deeper and facet by infancy_vac status
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()
```




```
# No obvious wP vs. aP differences even if focus on one visit
```