Pertussis Mini Project

Ari_Fon (PID: A15390446)

3/8/2022

Investigate Pertussis case numbers over time in the US

The CDC has tracked cases numbers since the early 1920s. https://www.cdc.gov/pertussis/surv-reporting/cases-by-year.html

```
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,
```

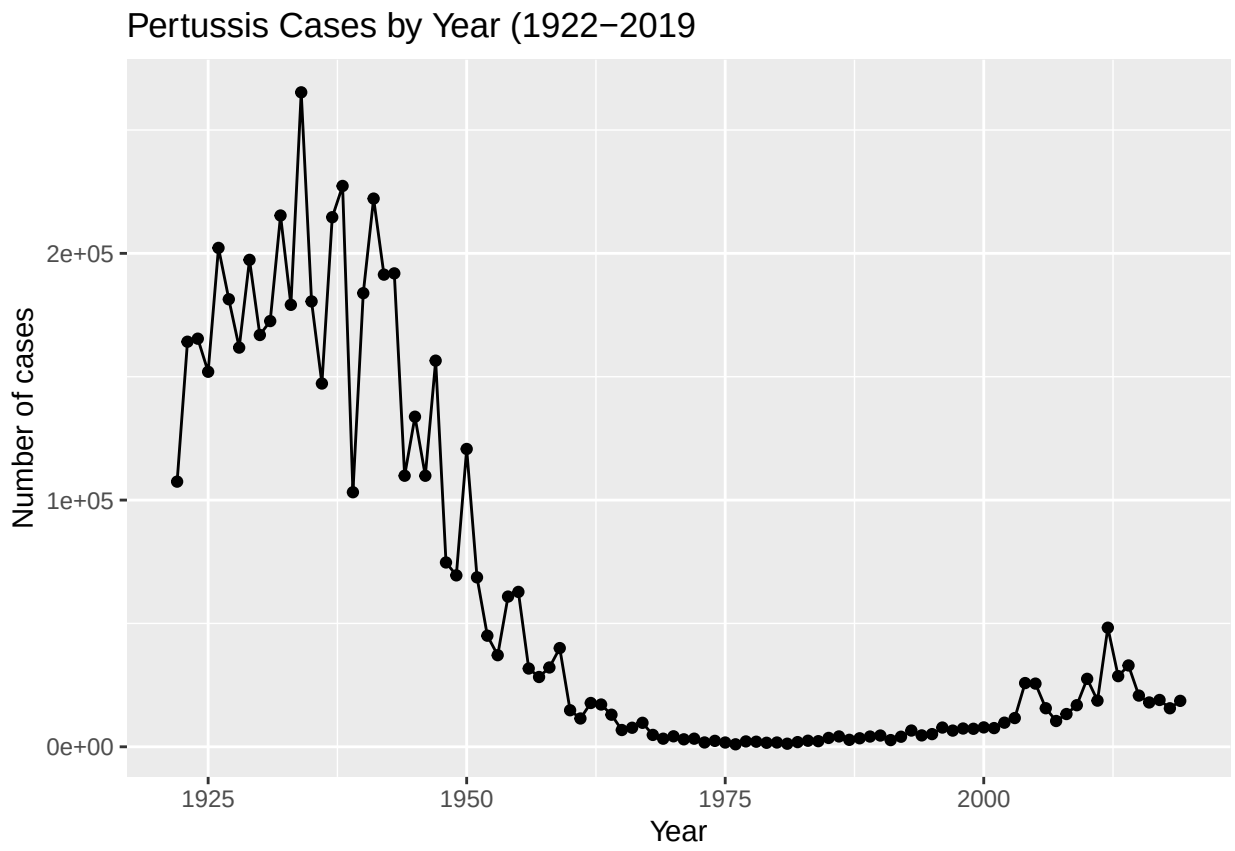
```
                                        16858,27550,18719,48277,28639,32971,
                                        20762,17972,18975,15609,18617)
       )
```

Q1. Make a figure of cases (y) per year (x) using ggplot

```
library(tidyverse)
```

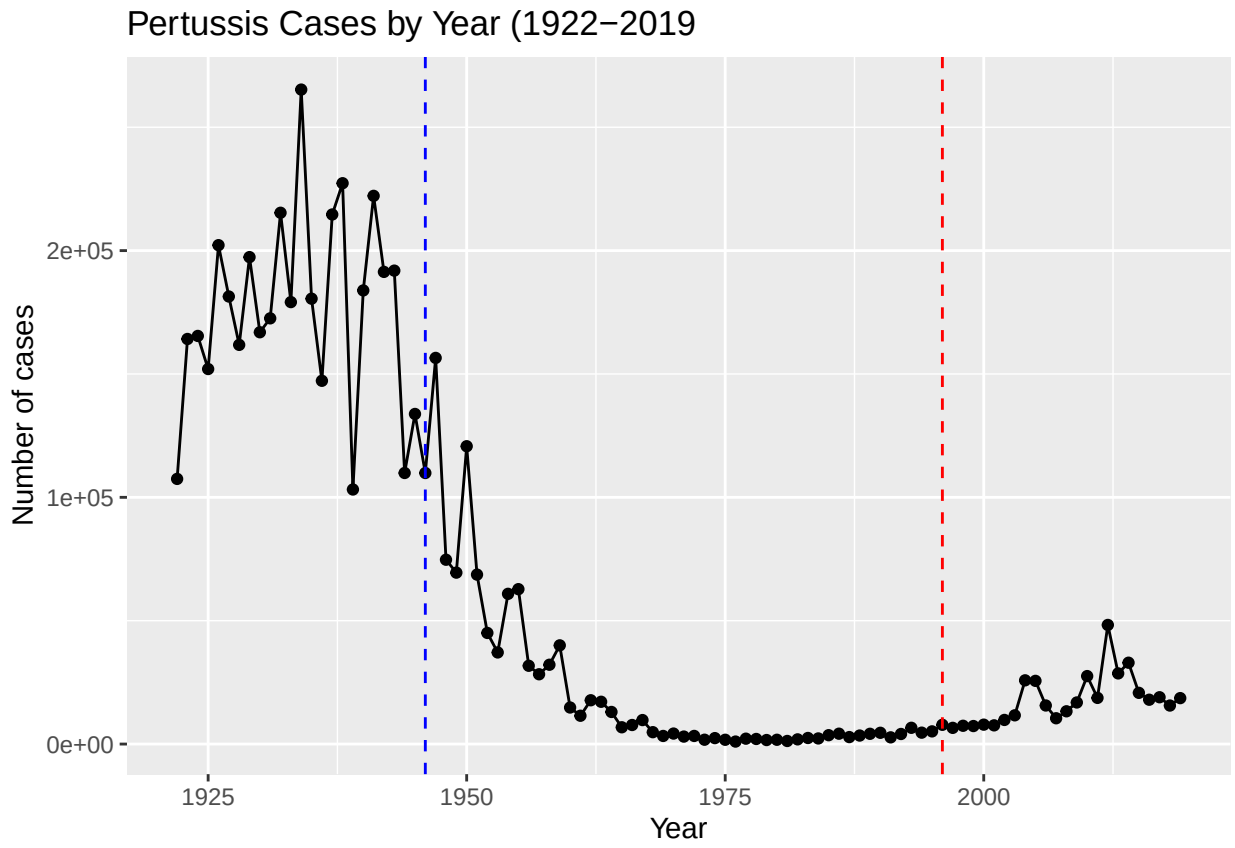
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## 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

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

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



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(title= "Pertussis Cases by Year (1922-2019", x= "Year", y= "Number of cases") +
  geom_vline(xintercept = 1946, color= "blue", linetype= "dashed") +
  geom_vline(xintercept = 1996, color= "red", linetype= "dashed")
```



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

After the 1946 vaccination the number of cases significantly dropped but after the 1996 vaccination the reported cases went up again. The bacteria could have mutated or evolved as it spread. The new vaccine may not have been as effective as the old one. More sensitive testing may also contributed so we can pick up more cases. Vaccine hesitancy may also have contributed as more people grew wary about vaccines. Vaccine efficacy decreases over times so it is possible that the new vaccine efficacy may decrease faster than the old one. Waning immunity in adolescents is a huge concern in communities.

Key-point: Despite high levels of acellular pertussis (aP) vaccination, the United States and other countries are now experiencing a significant resurgence in pertussis cases with large outbreaks now once again a major public health concern

Waning of immunity in adolescents originally primed as infants with the newer aP vaccine as compared to the older wP vaccine.

#Exploring CMI-PB data

We will use the **jsonlite** package to read from the CMI-PB database API directly.

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

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

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

```
url <- "https://www.cmi-pb.org/api/subject"
subject <- read_json(url, 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
```

How many total?

```
nrow(subject)
```

```
## [1] 96
```

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
```

Q7 and Q8 are optional.

#Joining Multiple Tables (datasets)

Read the specimen and ab_titer tables into R and store the data as specimen and titer named data frames.

```
# 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)
```

Have a quick look

```
head(specimen, 3)
```

```
##   specimen_id subject_id actual_day_relative_to_boost
## 1           1          1                           -3
## 2           2          1                          736
## 3           3          1                            1
##   planned_day_relative_to_boost specimen_type visit
## 1                             0         Blood     1
## 2                           736         Blood    10
## 3                             1         Blood     2
```

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:

I need to use inner_join() here.

```
library(dplyr)
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 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    19
```

```
head(abdata, 3)
```

```
##   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
##   lower_limit_of_detection subject_id actual_day_relative_to_boost
## 1                      NaN          1                           -3
## 2                    29.17          1                           -3
## 3                     0.53          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
##                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
```

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 is way smaller compared to the other specimens. This is because we are still collecting data which is why it is smaller. Visit is a proxy for time.

Examine IgG1 Ab titer levels

Now using our joined/merged/linked abdata dataset filter() for IgG1 isotype and exclude the small number of visit 8 entries.

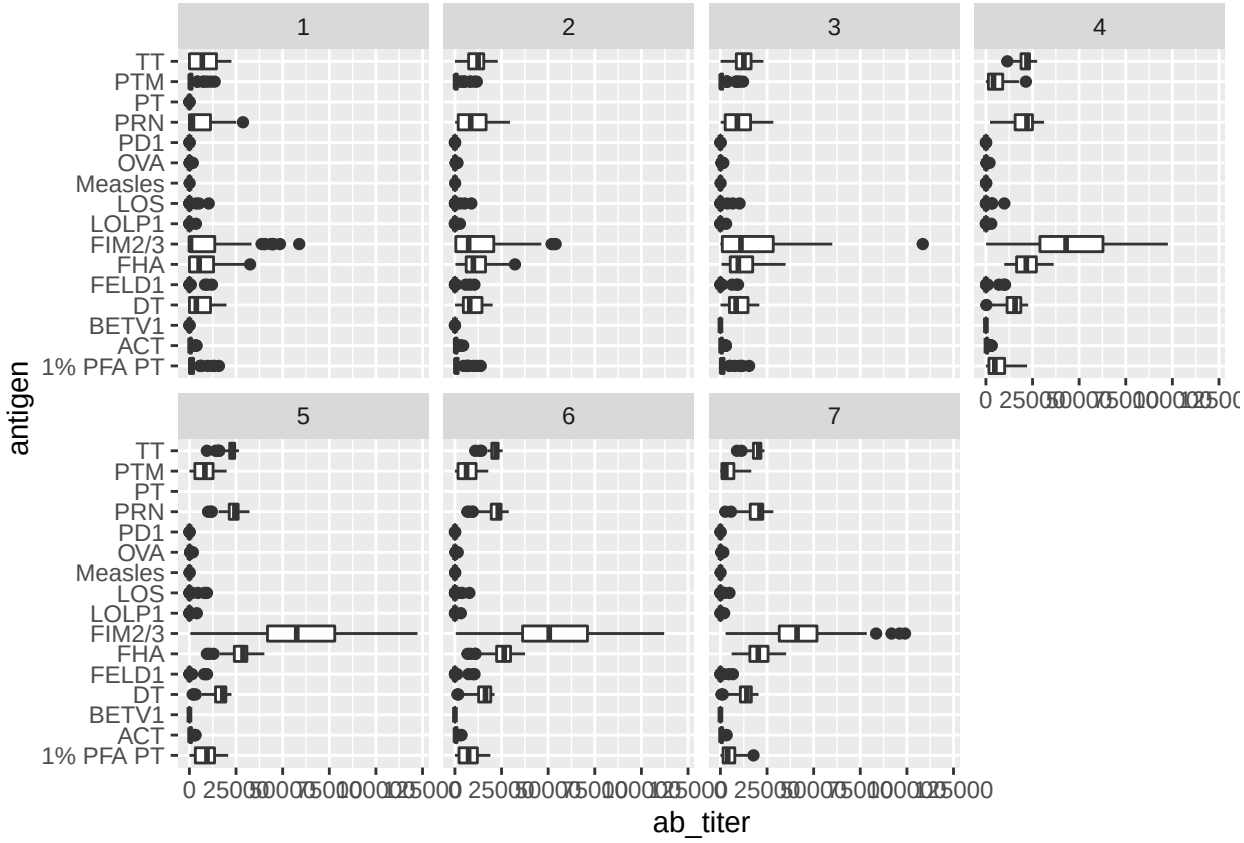
```
ig1 <- abdata %>% filter(isotype == "IgG1", visit!=8)
head(ig1)
```

```
##   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
```

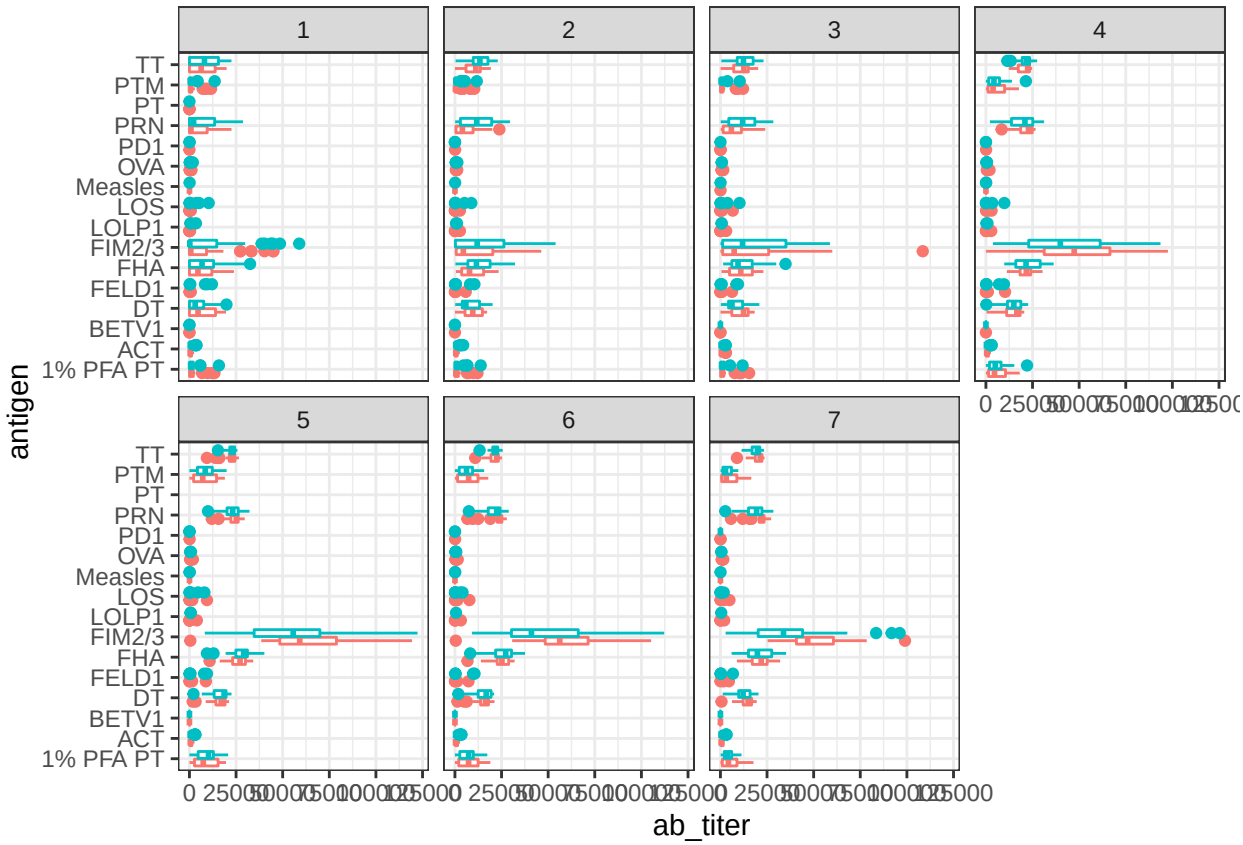
Q13. Complete the following code to make a summary boxplot of Ab titer levels for all antigens:

```
library(ggplot2)
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?

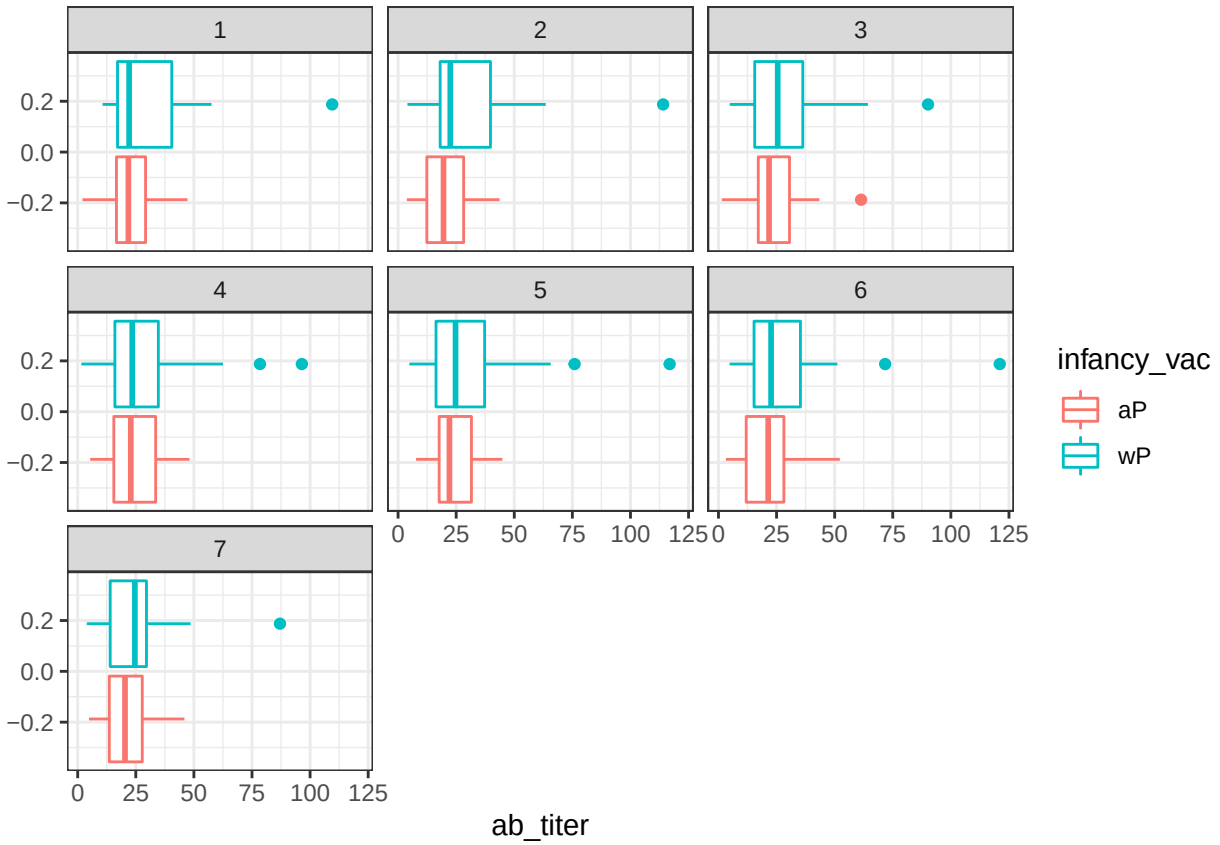
```
ggplot(ig1) +
  aes(ab_titer, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  theme_bw()
```



It seems that FIM2/3, TT, and PTM show differences in the level of IgG1 antibody titers recognizing them over time.This might be because antibodies specifically recognize them.

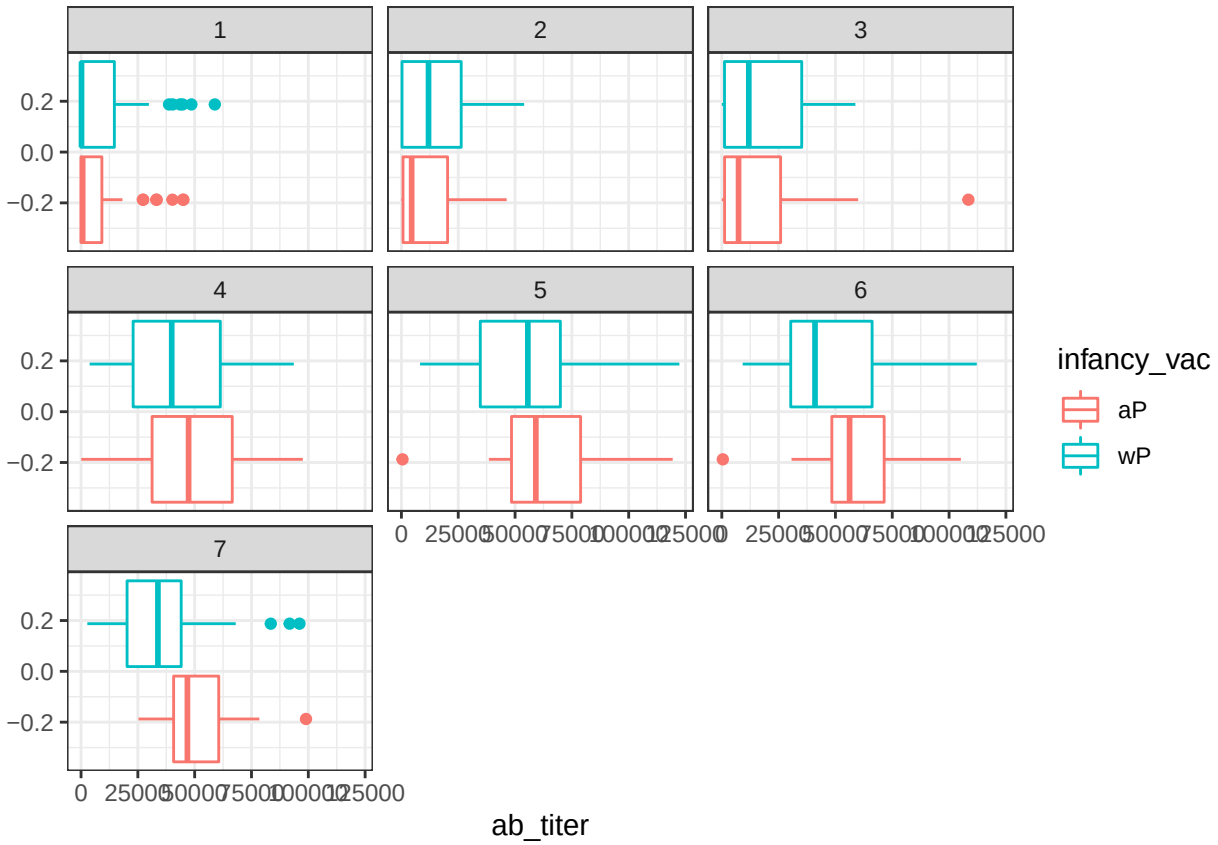
Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("Measles", that is not in our vaccines) and a clear antigen of interest ("FIM2/3", extra-cellular fimbriae proteins from B. pertussis that participate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



It appears to stay relatively consistent at low levels. This is the control after all.

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



FIM2/3 experienced an increase over time and this increase is around the same for both vaccines.In comparison to the control, FM2/3 increases and appears to peak around visit 4.

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

FIM2/3 in particular had a greater rise over the time course

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

There doesn't appear to be any clear differences in aP vs, wP respones. In FIM2/3 there was a slight difference between aP and wP but there is overlap so no significant difference.

Pull RNA-Seq data from the CMI-PB

We can use the CMI-PB API to obtain time-cours RNA seq results for wP and aP subjects (i.e. patients)

```
url2 <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.7"

rna <- read_json(url2, simplifyVector = TRUE)
dim(rna)
```

```
## [1] 360   4
```

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

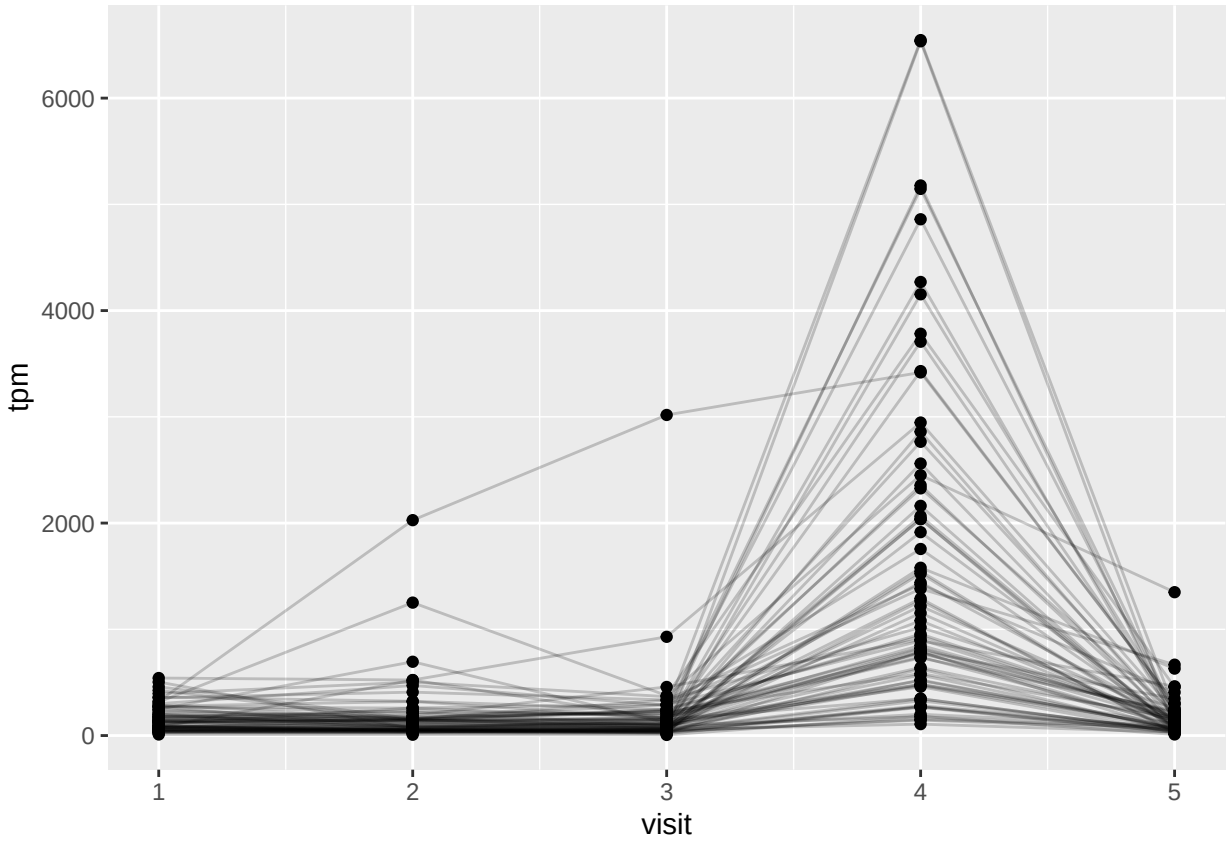
```
## Joining, by = "specimen_id"
```

```
dim(ssrna)
```

```
## [1] 360  16
```

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)
```



Notice: there is a rise at visit 4

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

It appears that the expression of this gene, i.e. when its at its maximum level, is a little above 6000 tpm at around visit 4.

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

It looks like this pattern in time matches the trend of the antibody titer data.