class15

Vince (PID: A15422556)

3/8/2022

Investigate Pertussis case numbers over time in the US

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

> 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.

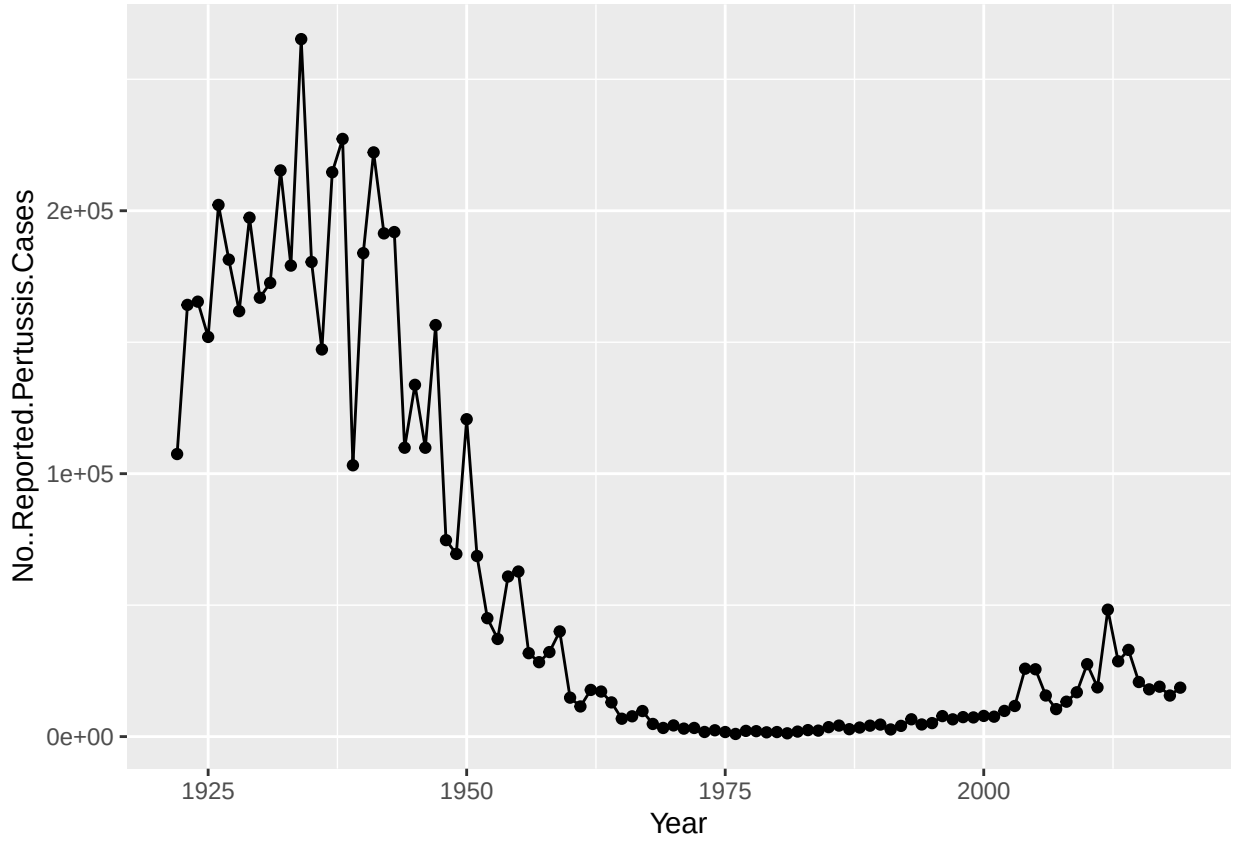
```
cdc <- data.frame(
                                     Year = c(1922L,1923L,1924L,1925L,
                                              1926L,1927L,1928L,1929L,1930L,1931L,
                                              1932L,1933L,1934L,1935L,1936L,
                                              1937L,1938L,1939L,1940L,1941L,1942L,
                                              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(tidyverse)
#ggplot2
```

```
plot <- ggplot(cdc) +
  aes(Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line()

plot
```

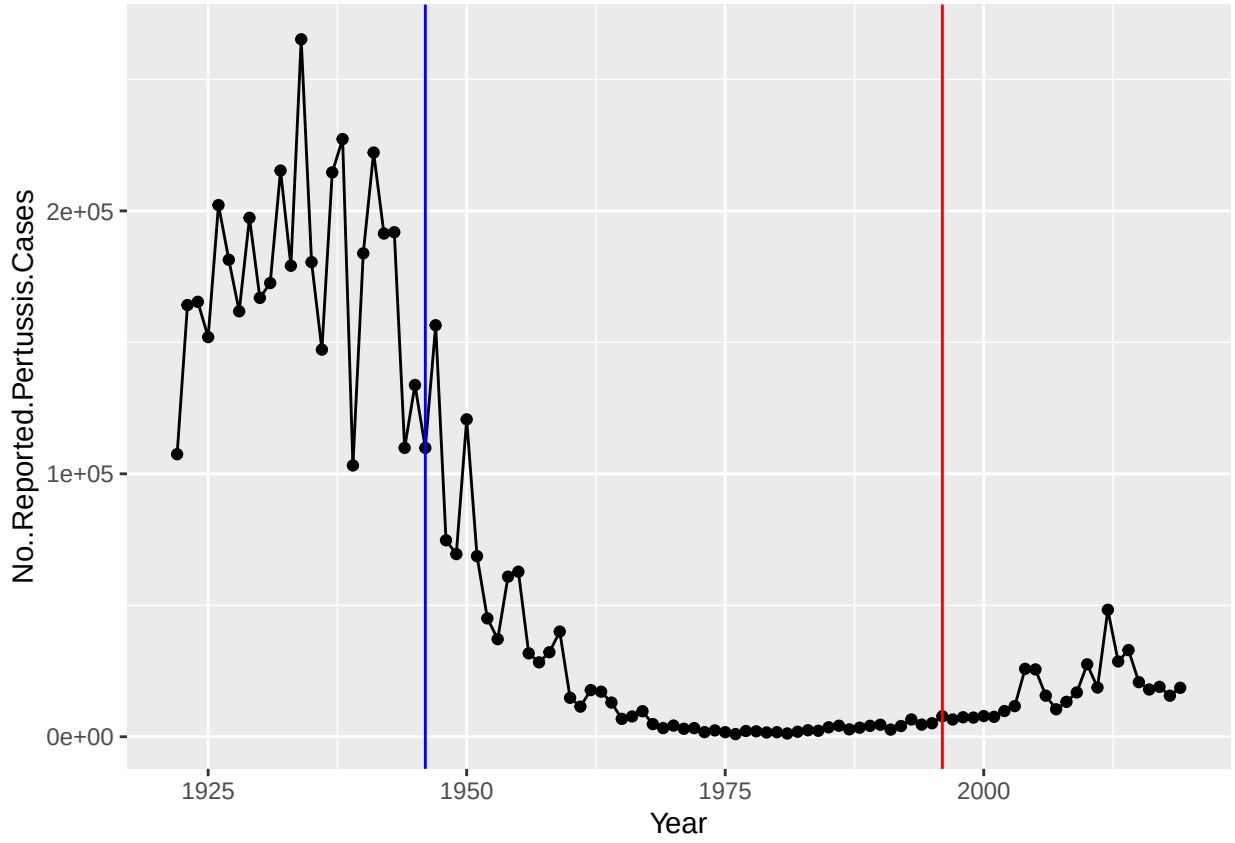


Q2. Using the ggplot geom_vline() function add lines to your previous plot for the 1946 introduction of the wP vaccine and the 1996 switch to aP vaccine (see example in the hint below). What do you notice? Introduction of the vaccine caused a drastic decline in Pertussis cases.

Q3. Describe what happened after the introduction of the aP vaccine? Do you have a possible explanation for the observed trend? A few years after introduction of the aP vaccine, there was

an increase in cases. This may be because of mutations in the Pertussis virus that renders the vaccine less effective than before. It is also possile the vaccine just became less effective over time and there is greater vaccine hesitancy.

```
plot +
  geom_vline(xintercept=1946, color="blue") +
  geom_vline(xintercept=1996, color="red")
```



CMI-PB Data

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

```
library(jsonlite)
```

```
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? 47 aP, 49 wP

Q5. How many Male and Female subjects/patients are in the dataset? 30 males, 66 females

Q6. What is the breakdown of race and biological sex (e.g. number of Asian females, White males etc...)? See code chunk below

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

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

```
table(subject$biological_sex)
```

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

```
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 optional.

Join 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. With the help of a faceted boxplot (see below), do you think these two groups are significantly different? Yes, these groups are significantly different based on their different distribution of values

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  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)
```

```
##   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
```

Q12. What do you notice about the number of visit 8 specimens compared to other visits? There are much less visit 8 specimens compared to other visits because the project is still ongoing

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

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

IgG1 Ab titer levels

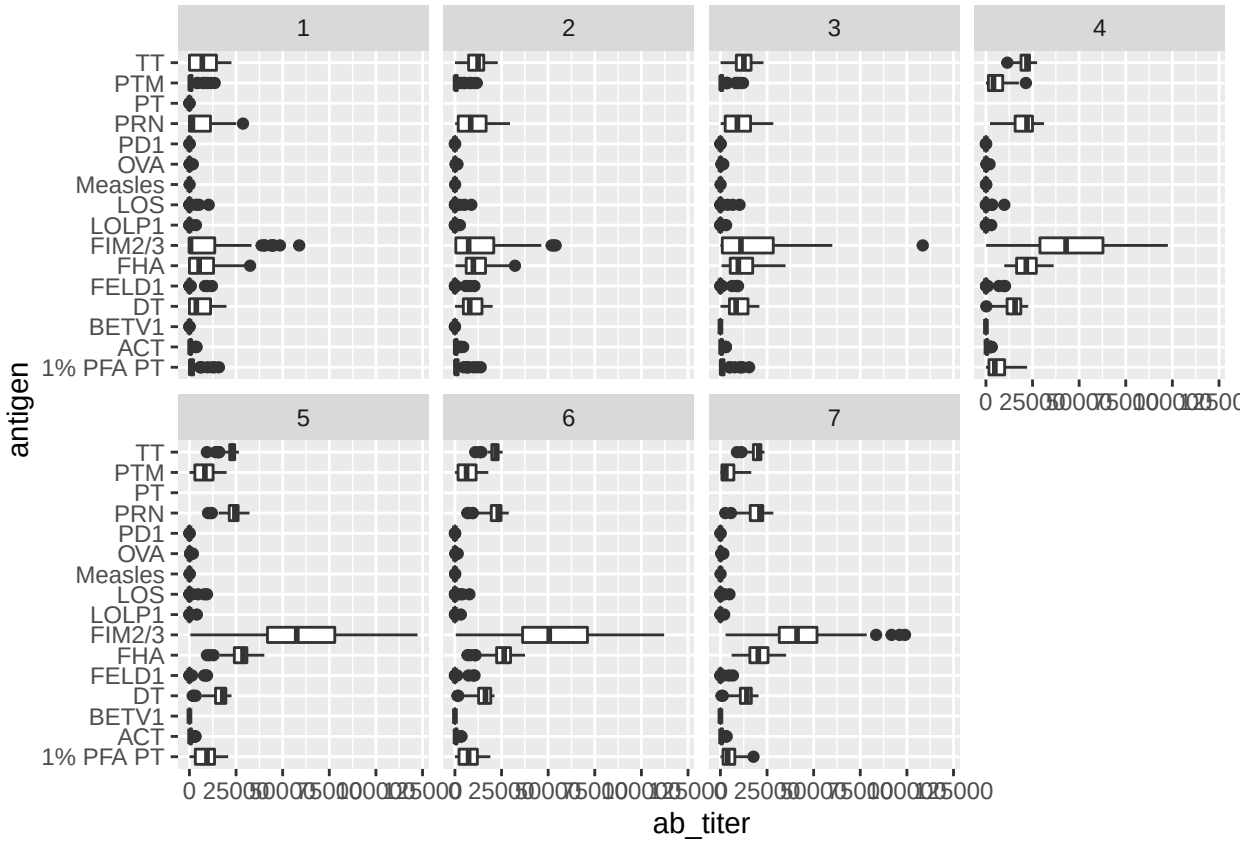
```
ig1 <- abdata %>% filter(isotype == "IgG1", visit!=8)
head(ig1)
```

```
##   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
## 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
```

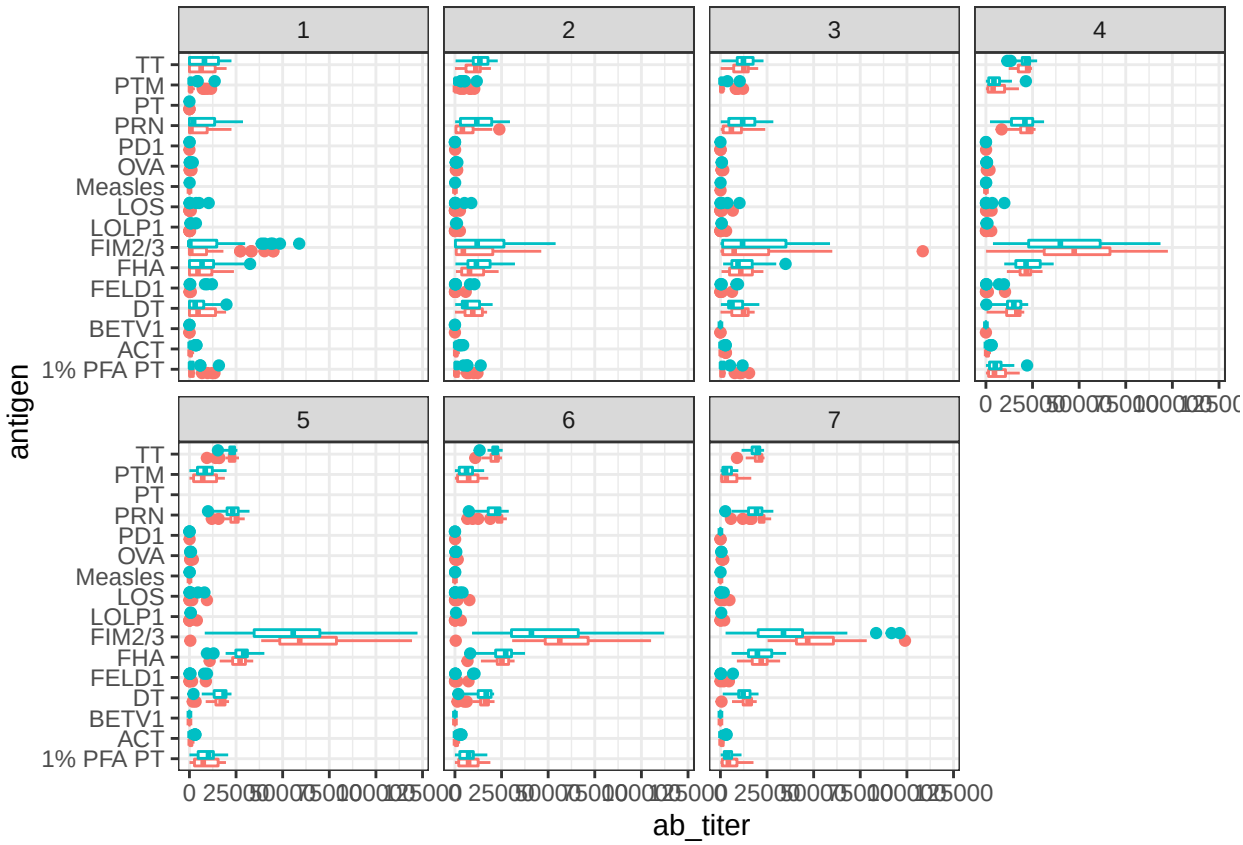
Q13. Complete the following code to make a summary boxplot of Ab titer levels for all antigens:

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

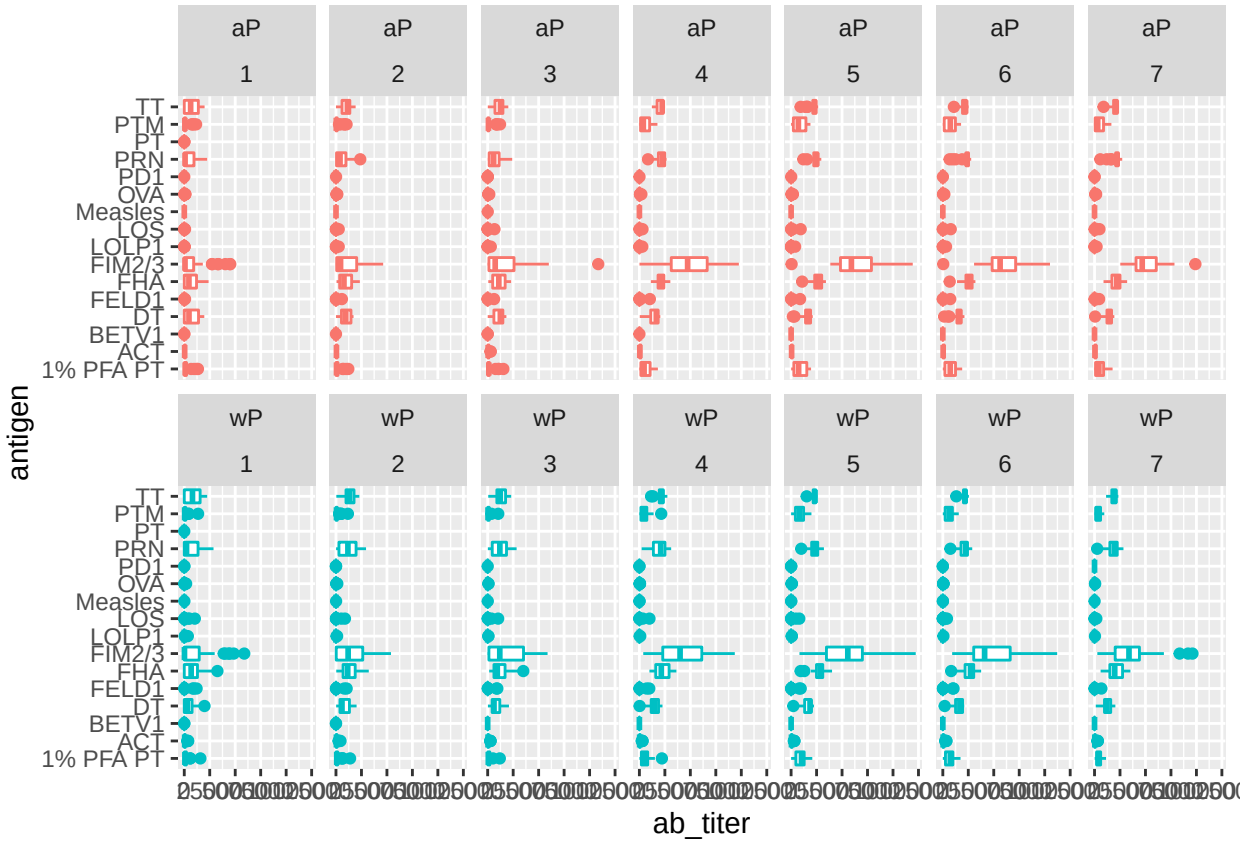


Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others? FIM2/3 and FHA increases over time because there are antibodies for these specific antigens in the vaccine (fimbriae); results in an antigen response

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

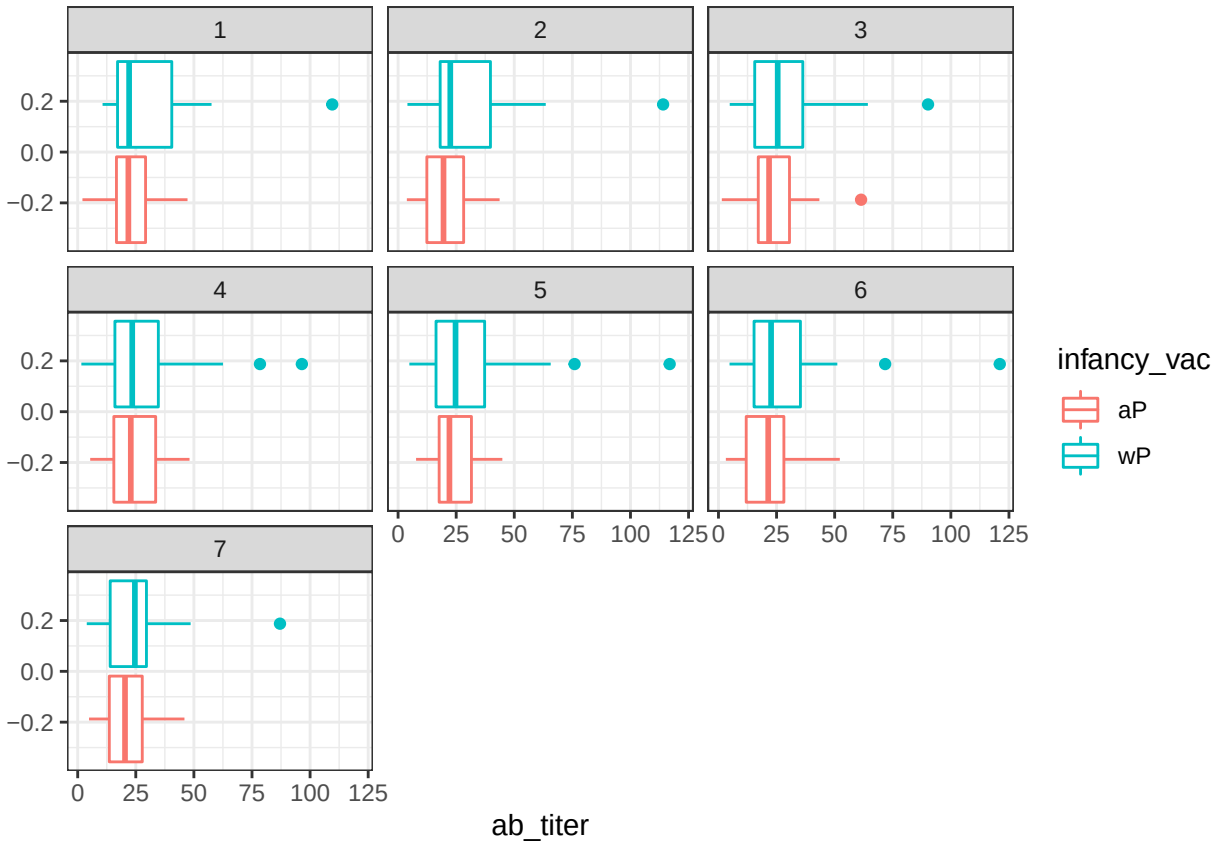



```
ggplot(ig1) +
  aes(ab_titer, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```



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).

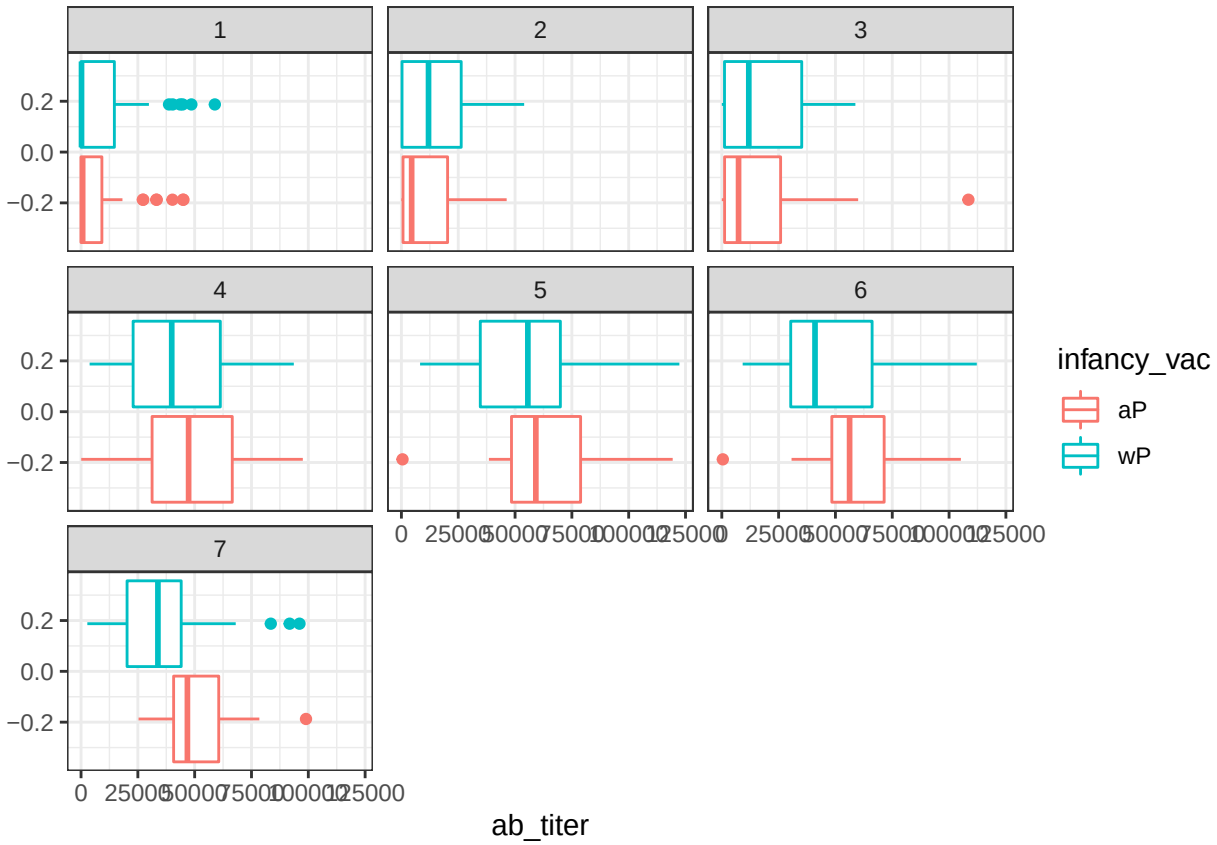
```
filter(ig1, antigen=="Measles") %>%
  ggplot() +
  aes(ab_titer, col=infancy_vac) +
  geom_boxplot(show.legend = TRUE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



Q16. What do you notice about these two antigens time course and the FIM2/3 data in particular? The measles Ab titer data generates little to no antigen response, while the FIM2/3 data produces a strong antigen response

Q17. Do you see any clear difference in aP vs. wP responses? No clear differences

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



CMI-PB RNASeq data

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

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

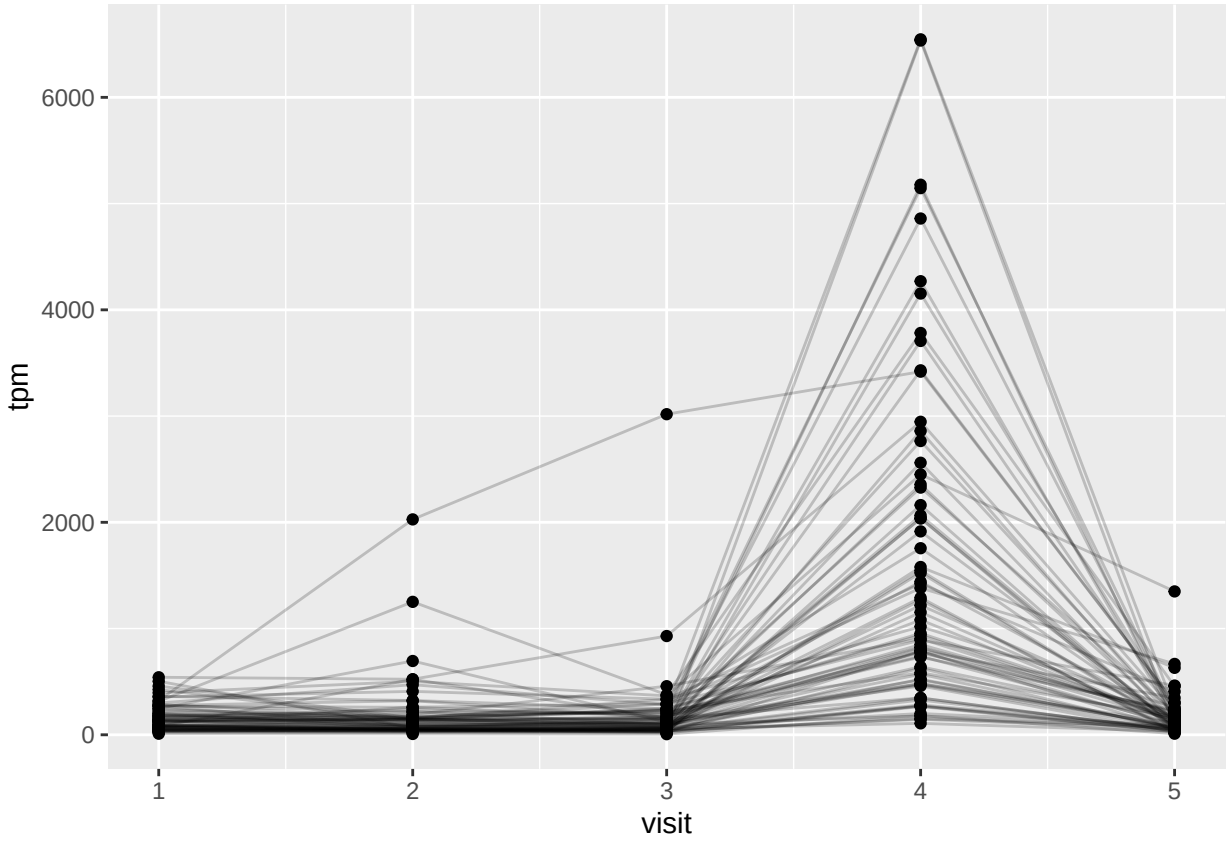
```
## [1] 360   4
```

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

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

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)? The gene is maximally expressed at visit 4

Q20. Does this pattern in time match the trend of antibody titer data? If not, why not? Yes this matches the trend of Ab titer data since at visit 4 there is a spike in the antigen response; there is lag as the spike in antigen response is not until visit 5 which is likely due to the fact that gene expression is ahead of antigen response