

Project15

Haroon Riyaz (PID A15377799)

3/14/2022

1. Investigating Pertussis Bases Each 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(  
  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 (1922-2019)", x = "Year", y = "Number of Cases")
```



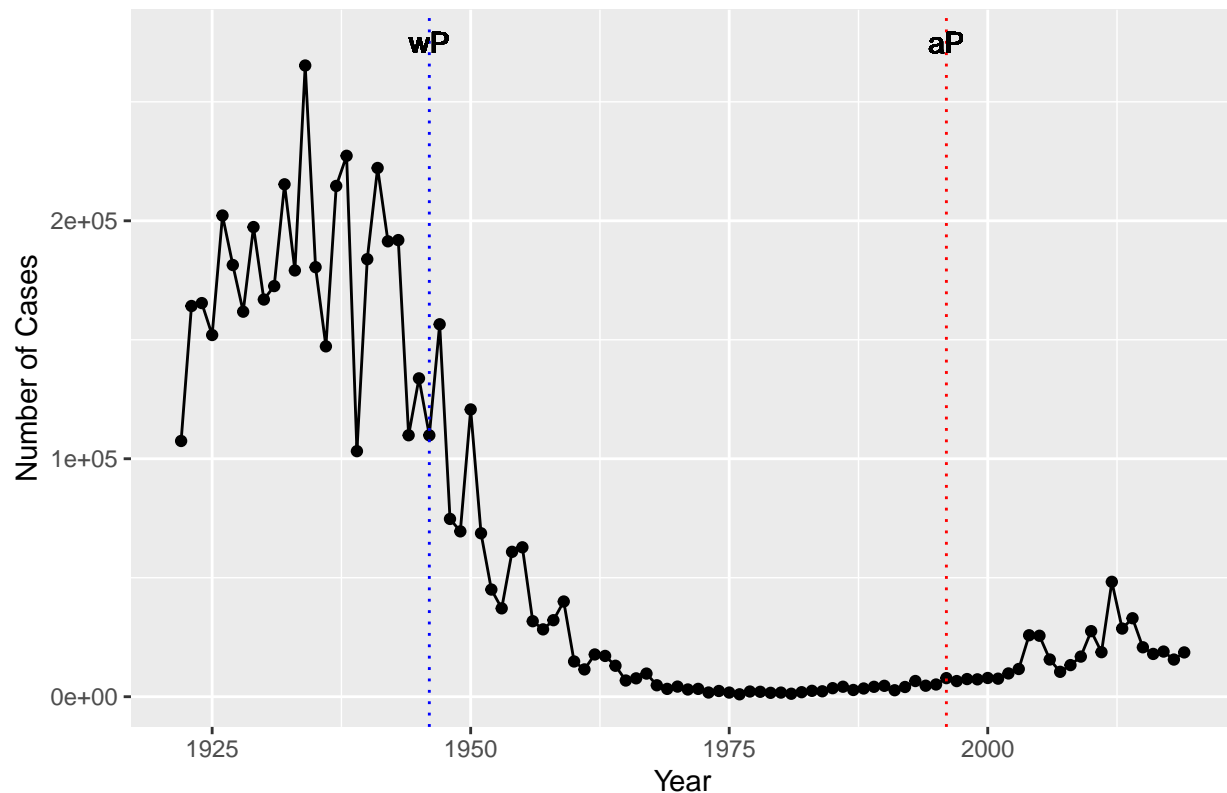
2. A tale of two vaccines (wP & aP)

Switch-over to New Vaccine

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(x
  geom_vline(xintercept = 1996L, linetype = "dotted", color = "red") + geom_text(aes(1946L, 275000), la
```

Pertussis Cases by Year (1922–2019)



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

There is an increase in the number of cases of Pertussis with nearly 50,000 cases around 2012. The increase in the number of cases could be due to hysteria/mistrust of vaccines leading to less people taking it, resulting in more cases. It could also be other factors like more effective testing, decreased immunity in adults who took the vaccine earlier on in life, and evolution of the bacteria.

3. Exploring CMI-PB data

Track immune response data of many people taking the vaccines

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

```
subject <- read_json("https://www.cmi-pb.org/api/subject", simplifyVector = TRUE)
```

```
# Example of data
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?

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

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

47 aP vaccinated subjects and 49 wP vaccinated subjects

Q5. How many Male and Female subjects/patients are in the dataset?

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

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

66 female subjects and 30 male subjects

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
```

Working with Dates

```
# Makes processing of dates easier
library(lubridate)
```

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

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

```
today()
```

```
## [1] "2022-03-14"
```

```
# Days passed since new year 2000
today() - ymd("2000-01-01")
```

```
## Time difference of 8108 days
```

```
# Length in years
time_length(today() - ymd("2000-01-01"), "years")
```

```
## [1] 22.19849
```

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?

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

```
#aP
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 <- subject %>% filter(infancy_vac == "aP")
round( summary( time_length(ap$age, "years") ) )
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.   Max.
##       22      24      25      25      25      26
```

```
#wP
library(dplyr)
wp <- subject %>% filter(infancy_vac == "wP")
round( summary( time_length(wp$age, "years") ) )
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       22      24      25      25      25      26
```

Age ranges, medians, means and other age statistics are significantly different.

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

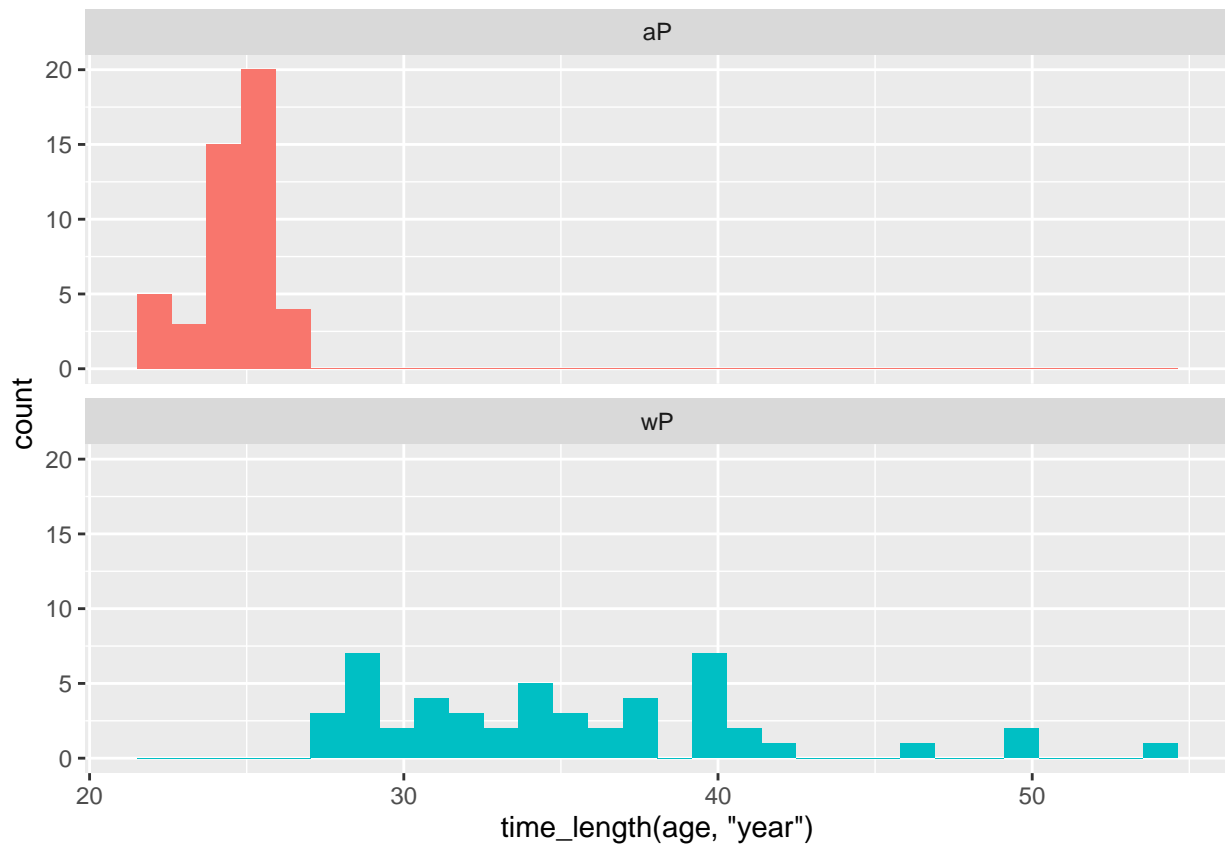
```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

```
## [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?

```
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'.
```



The two groups appear to be significantly different since nearly all the aP population was much younger at time of booster shot while the wP population took the booster shot at a wide range of ages and later in life.

Joining multiple tables

```
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)
```

Q10. 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:

```
library(dplyr)
meta <- left_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
## 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
## 1 13221 days
## 2 13221 days
## 3 13221 days
## 4 13221 days
## 5 13221 days
## 6 13221 days
```

Q11. 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
```

Q12. 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
```

Q13. 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 are a lot less visits for the #8 specimens

4. Examine IgG1 Ab titer levels

```
# Exclude # of Visit 8 Entries
ig1 <- abdata %>% filter(isotype == "IgG1", visit!=8)
head(ig1)
```

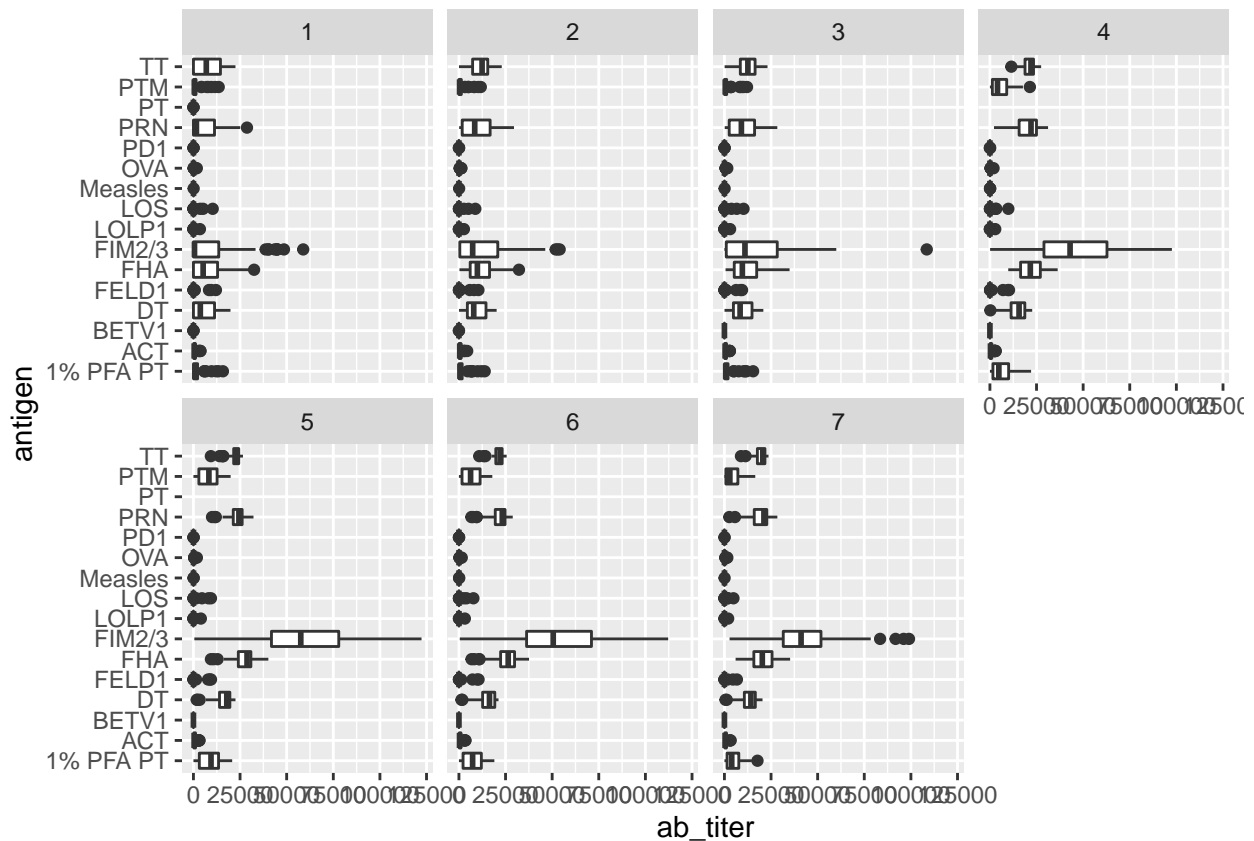
```
## specimen_id isotype is_antigen_specific antigen  ab_titer  unit
## 1           1   IgG1                TRUE    ACT 274.355068 IU/ML
## 2           1   IgG1                TRUE    LOS 10.974026 IU/ML
## 3           1   IgG1                TRUE  FELD1  1.448796 IU/ML
## 4           1   IgG1                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
##          age
## 1 13221 days
## 2 13221 days
## 3 13221 days
## 4 13221 days
## 5 13221 days
## 6 13221 days
```

Q14. Complete the following code to make a summary boxplot of Ab titer levels for all antigens:

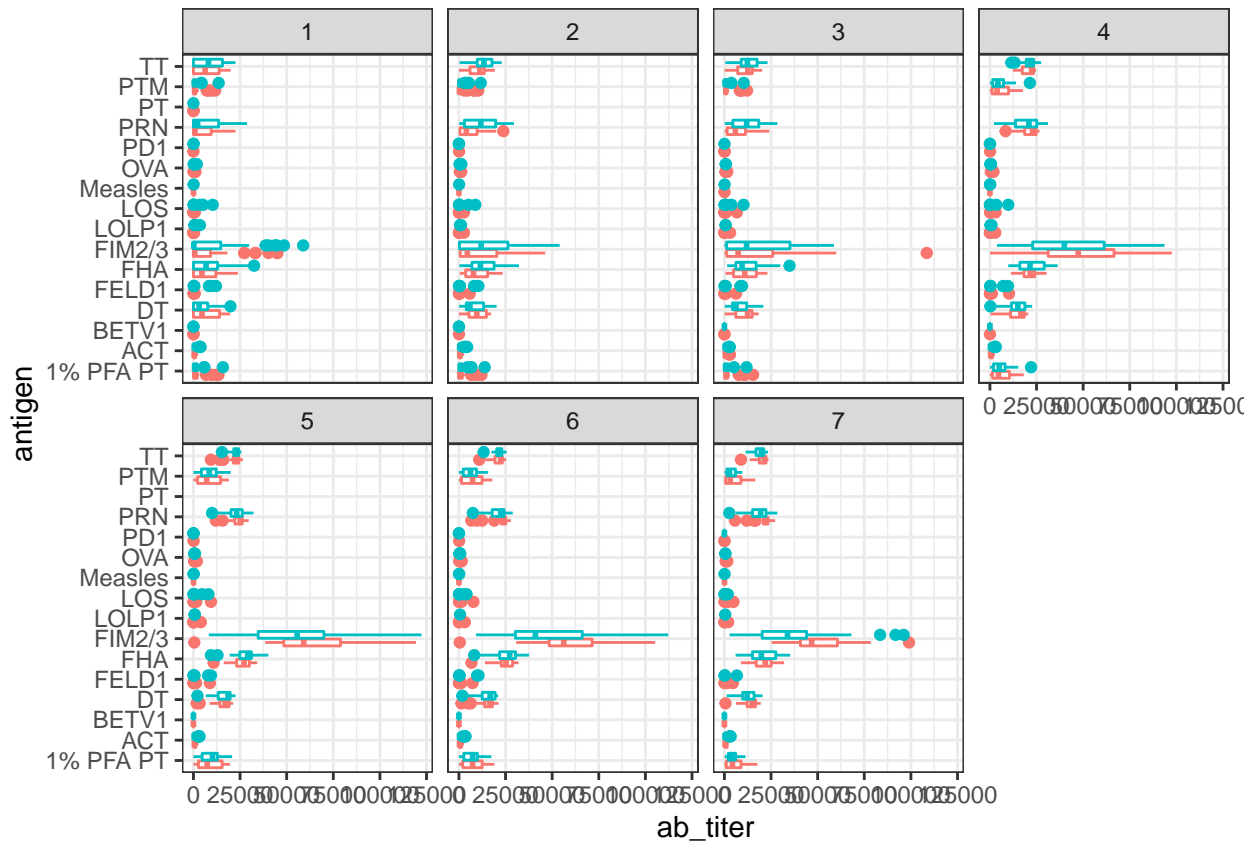
```
ggplot(ig1) +
  aes(ab_titer, antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow=2)
```



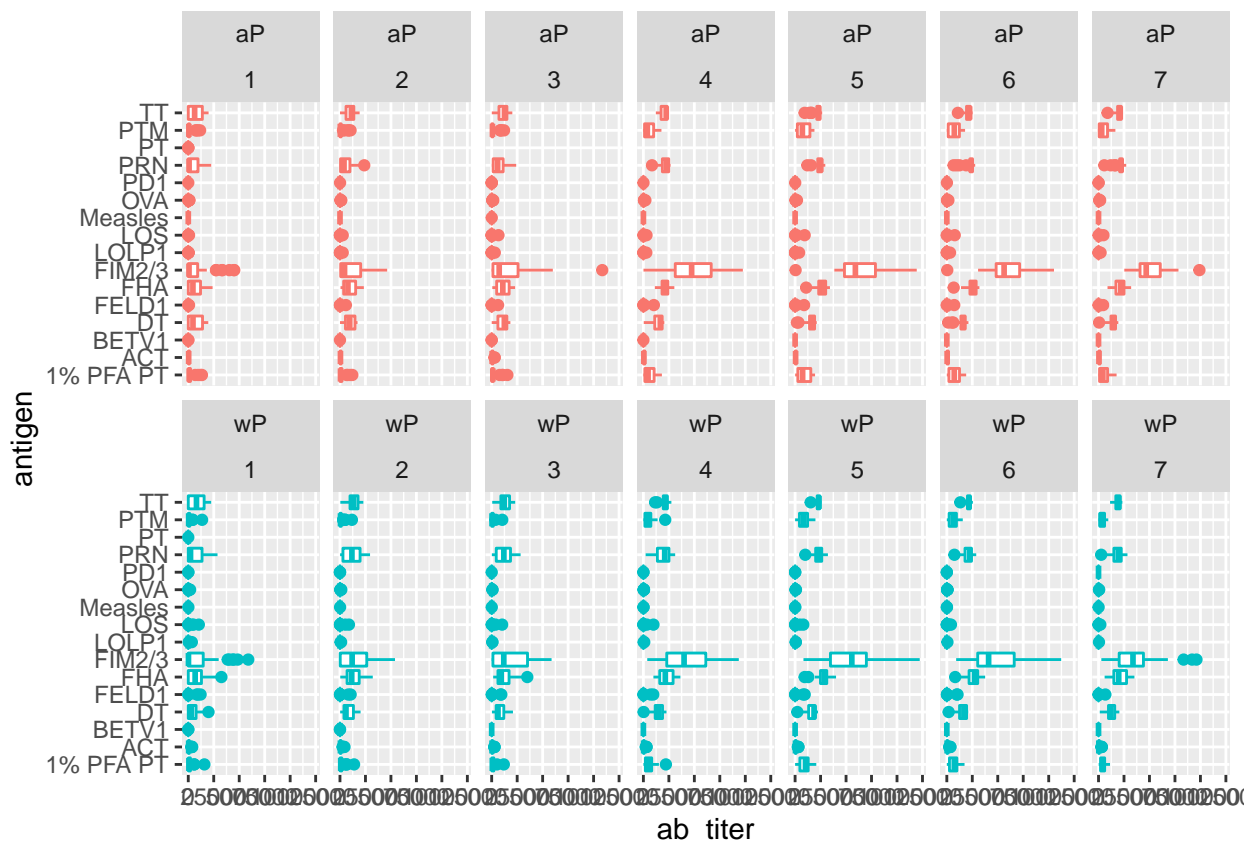
> Q15. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

The FIM2/3 antigen shows differences in the level of IgG1 antibody recognizing them over time. IgG1 most likely recognizes these and not others since the antigen belongs to Bordetella Pertussis and will thus be recognized by the immune system, hence why there is increased antibody titers.

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



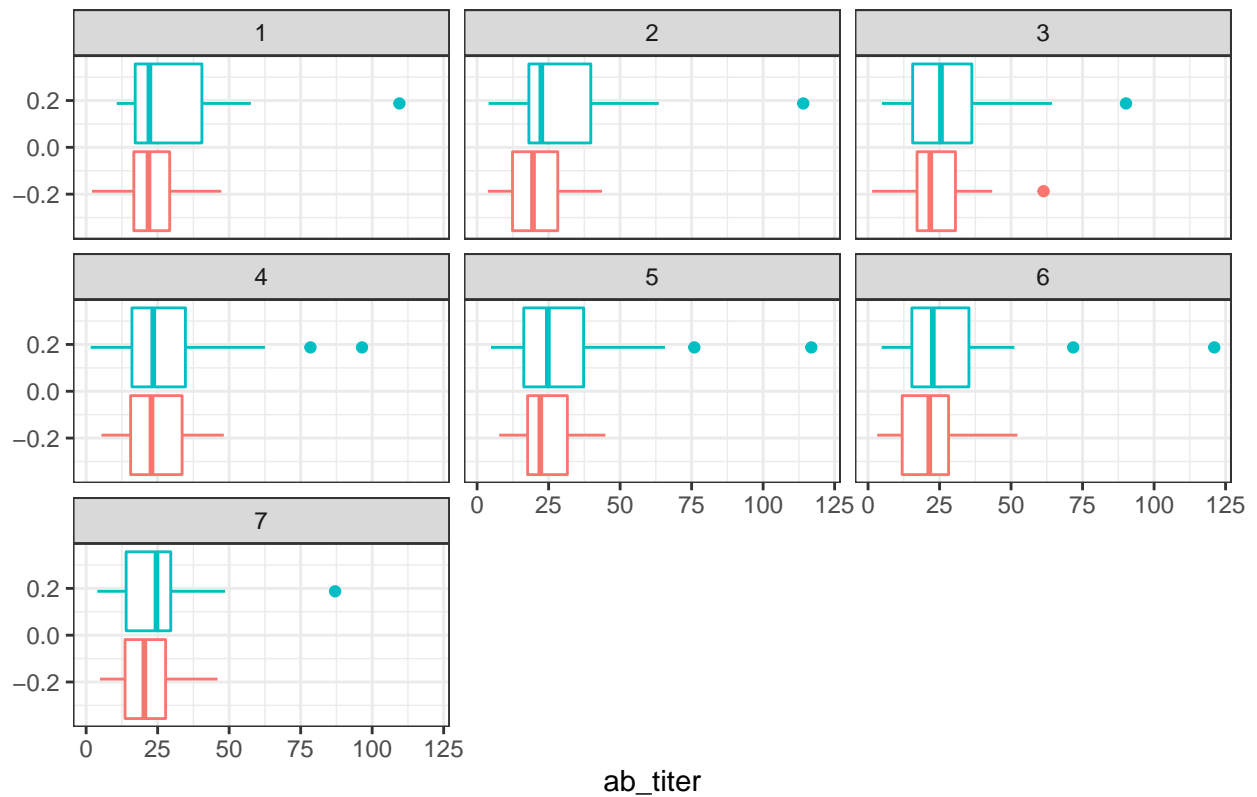
```
# infancy_vac included
ggplot(ig1) +
  aes(ab_titer, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(infancy_vac, visit), nrow = 2)
```



> Q16. 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).

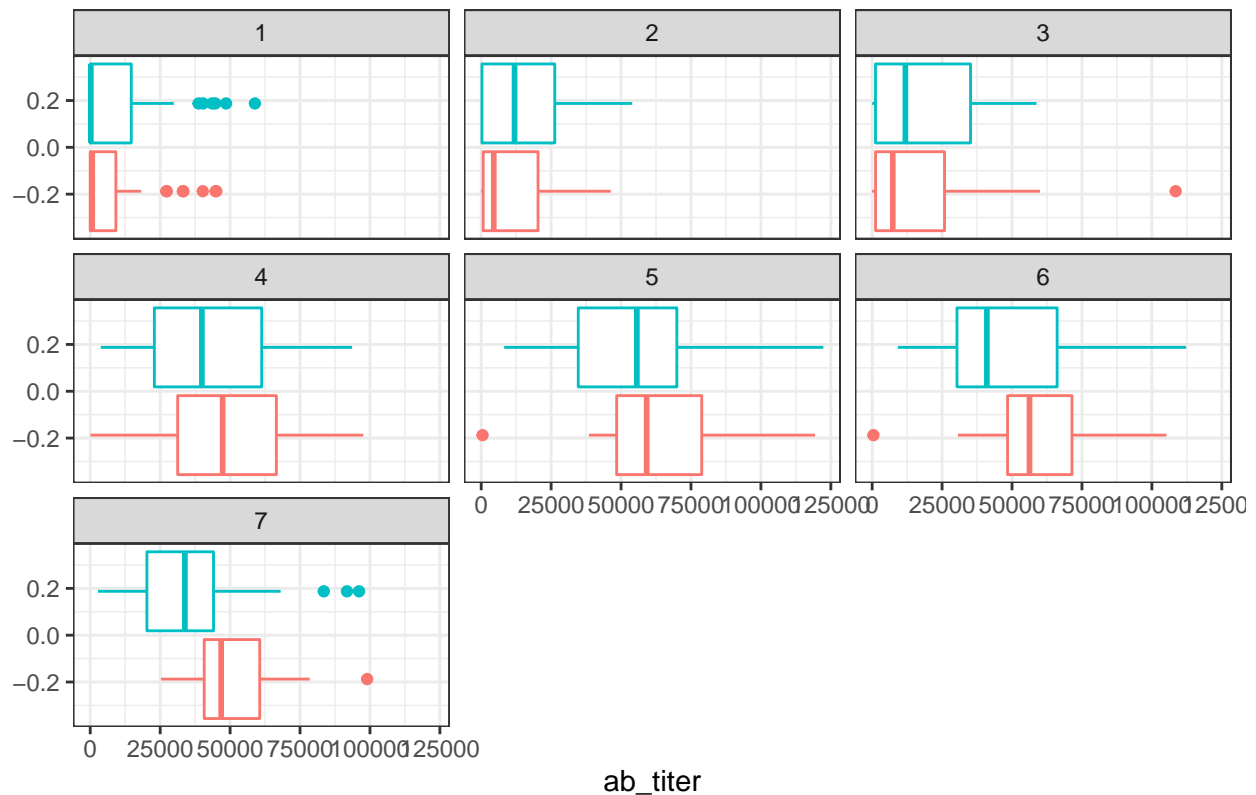
```
filter(ig1, antigen == "Measles") %>%
  ggplot() +
  aes(ab_titer, col = infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw() + labs(title = "Measles Antigen Levels Per Visit (aP = red, wP = teal)")
```

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)")
```

FIM2/3 Antigen Levels Per Visit (aP = red, wP = teal)



> Q17. What do you notice about these two antigens time course and the FIM2/3 data in particular?

For both wP and aP, the FIM2/3 ab_titer levels rise and go far above the level of measles which makes sense the antibody will not detect measles but instead FIM2/3. FIM2/3 ab_titer levels are the highest at visit 5 and then decline.

Q18. Do you see any clear difference in aP vs. wP responses?

The FIM2/3 ab_titer levels for the aP population are slightly higher than that of the wP population.

5. Obtaining CMI-PB RNASeq data

```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENS00000211896.7"
rna <- read_json(url, simplifyVector = TRUE)
```

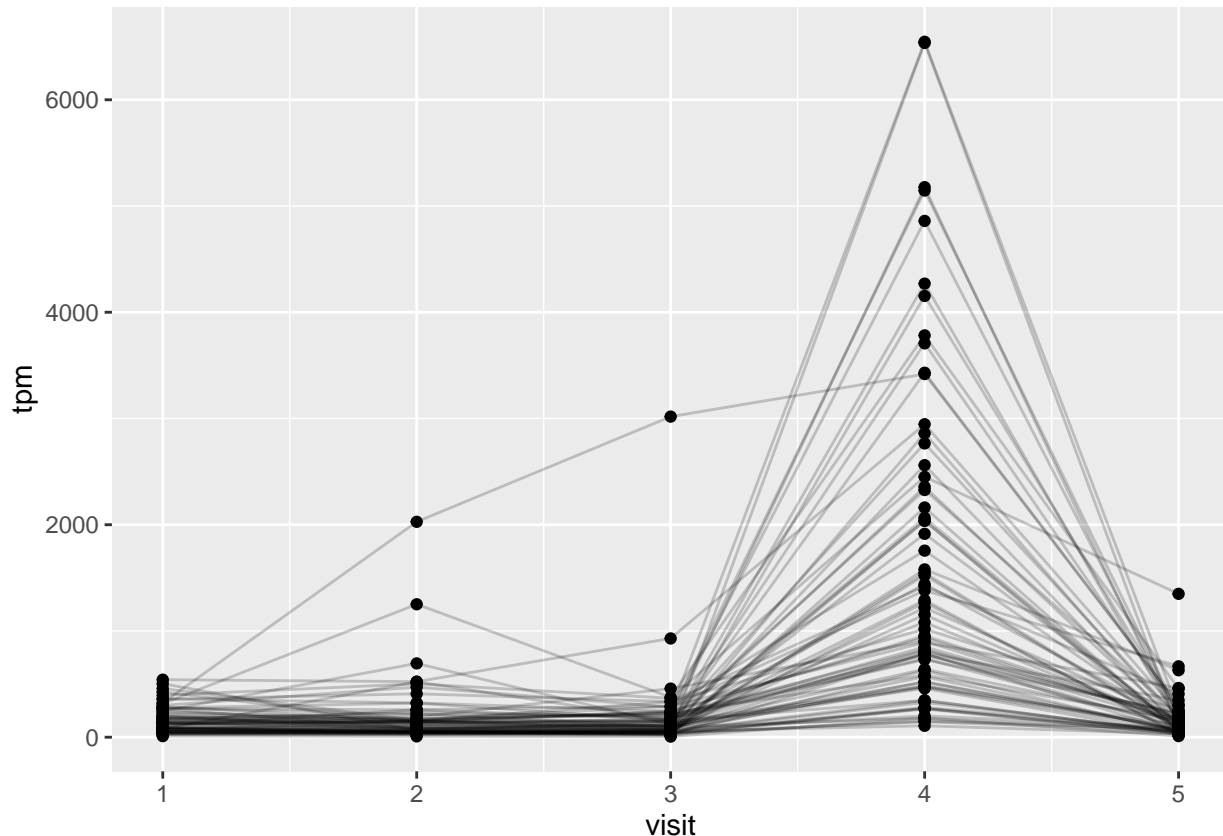
Look at gene expression of IGHG1 gene with aP/wP and # of visits

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

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

Q19. 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)
```



> Q20.: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

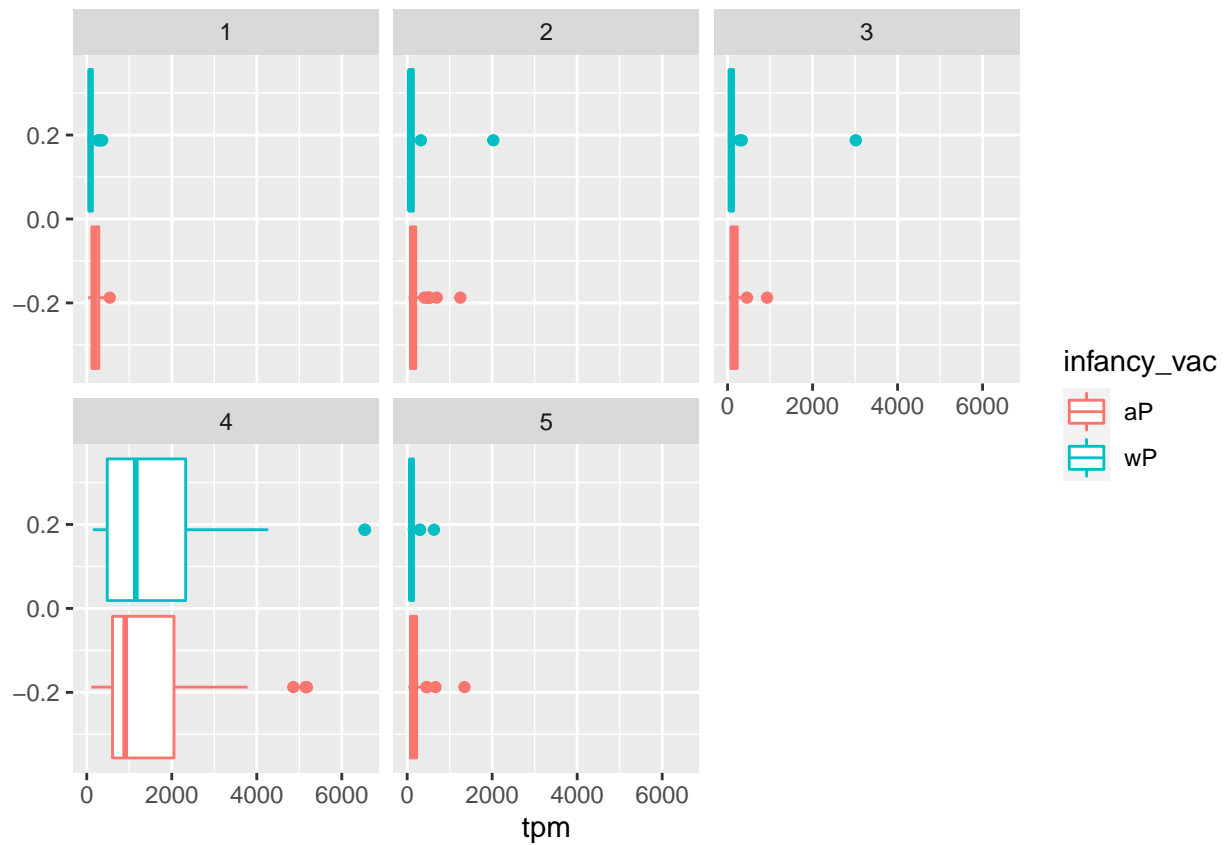
Expression is at the maximum level at the 4th visit. Expression increases until this 4th visit where it is at the maximum. Expression then declines during the 5th visit.

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

Yes, *but not exactly*. Both antibody titer and tpm data show an increase in value as the number of visits increases until an intermediate number of visits (4-5) after which it declines in value (5-7).

HOWEVER, the antibody titer value is at the maximum after 5 visits compared to the tpm value which is at the maximum after 4 visits. This is because antibodies last longer than the cells which make them. Cells will not be able to last and thus the maximum production of antibody will take place at fewer visits. The antibodies last a long time and will be able to respond for much longer which is why the maximum is at 5 visits.

```
# Facet by infancy_vac
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()
```

