Pertussis Mini Project

Ethan Lai

3/9/2022

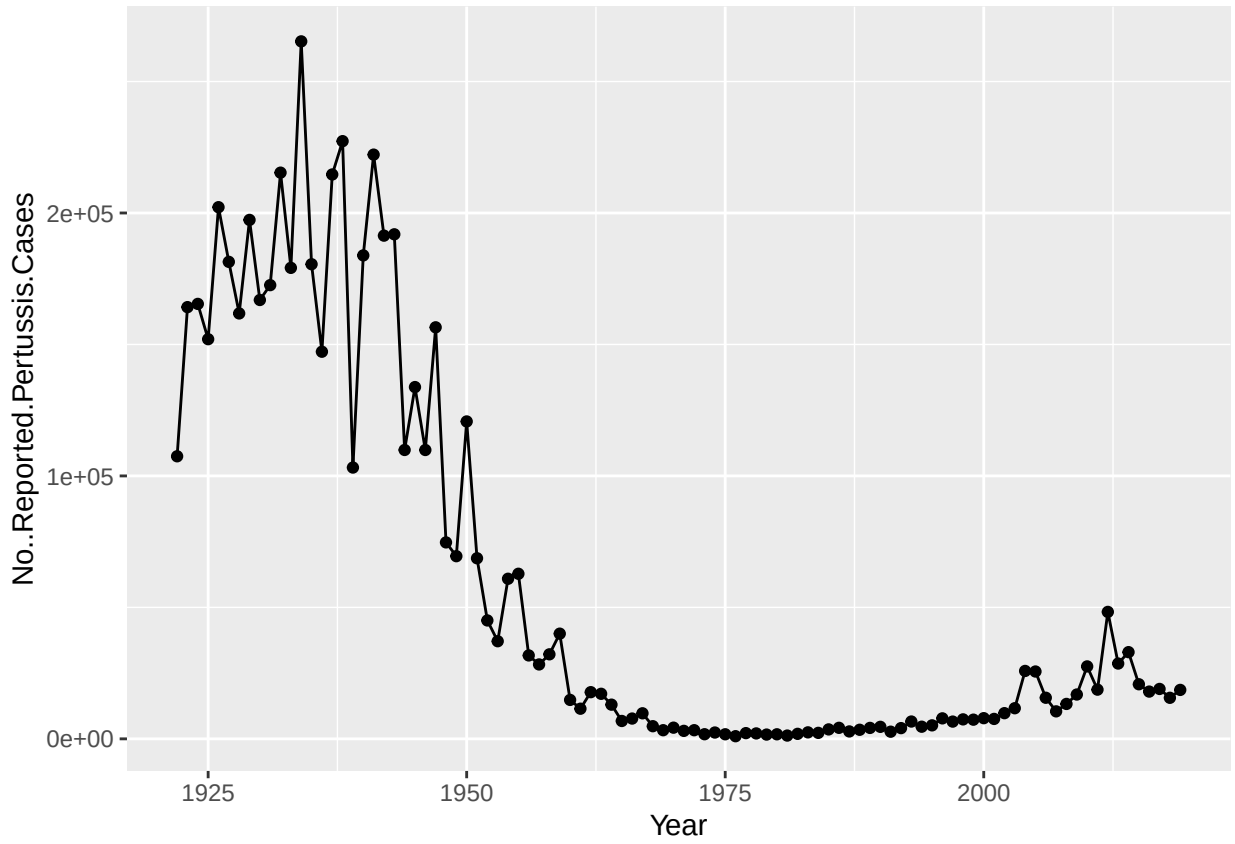
Q1. With the help of the R "addin" package datapasta assign the CDC pertussis case number data to a data frame called cdc and use ggplot to make a plot of cases numbers over time.

```
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)
    )
```

```
colnames(cdc)
```

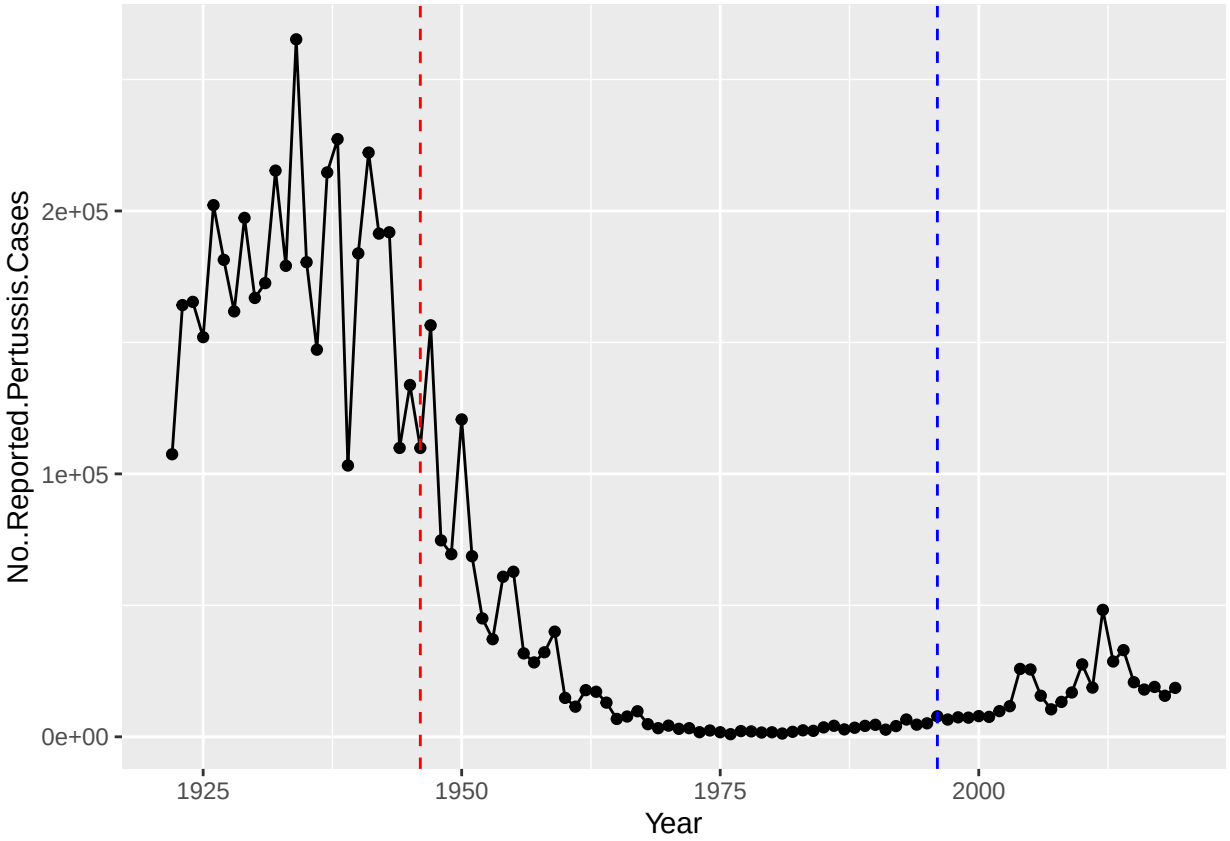
```
## [1] "Year"                          "No..Reported.Pertussis.Cases"
```

```
library(ggplot2)
cdcplot<- ggplot(cdc,aes(x=Year, y=No..Reported.Pertussis.Cases)) + geom_point() + geom_line()
cdcplot
```



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?

```
vaxYears <- c(1946, 1996)
cdcplot + geom_vline(xintercept = vaxYears, linetype="dashed", col=c("red", "blue"))
```



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

There seems to be a mild resurgence in pertussis rates. Maybe the aP vaccine is somewhat less effective? Or anti-vaxxers emerged around then...

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

```
##   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
## 4          4          wP           Male Not Hispanic or Latino Asian
## 5          5          wP           Male Not Hispanic or Latino Asian
## 6          6          wP         Female Not Hispanic or Latino 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
## 4    1988-01-01    2016-08-29 2020_dataset
## 5    1991-01-01    2016-08-29 2020_dataset
## 6    1988-01-01    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
```

49 wP subjects, 47 aP subjects.

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

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

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

66 female subjects, 30 male subjects

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

```
table(subject$race)
```

```
##
##             American Indian/Alaska Native
##                                         1
##                                     Asian
##                                        27
##                 Black or African American
##                                         2
##                           More Than One Race
##                                        10
## Native Hawaiian or Other Pacific Islander
##                                         2
##                   Unknown or Not Reported
##                                        14
##                                     White
##                                        40
```

```
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)

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
```

Now to make sense of this data and ask questions about ap vs wP of the Ab titer data. We need to join subject with these new tables.

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

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

```
dim(meta)
```

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

```
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 infancy_vac biological_sex
## 1                      NaN          1          wP         Female
## 2                29.170000          1          wP         Female
## 3                 0.530000          1          wP         Female
## 4                 1.070000          1          wP         Female
## 5                 0.064000          1          wP         Female
## 6                 2.816431          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
##   actual_day_relative_to_boost planned_day_relative_to_boost specimen_type
## 1                           -3                             0         Blood
## 2                           -3                             0         Blood
## 3                           -3                             0         Blood
```

```
## 4                           -3                           0         Blood
## 5                           -3                           0         Blood
## 6                           -3                           0         Blood
##   visit
## 1     1
## 2     1
## 3     1
## 4     1
## 5     1
## 6     1
```

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
```

There are far, far fewer visit 8s compared to the others. So, let's filter it out for later analysis

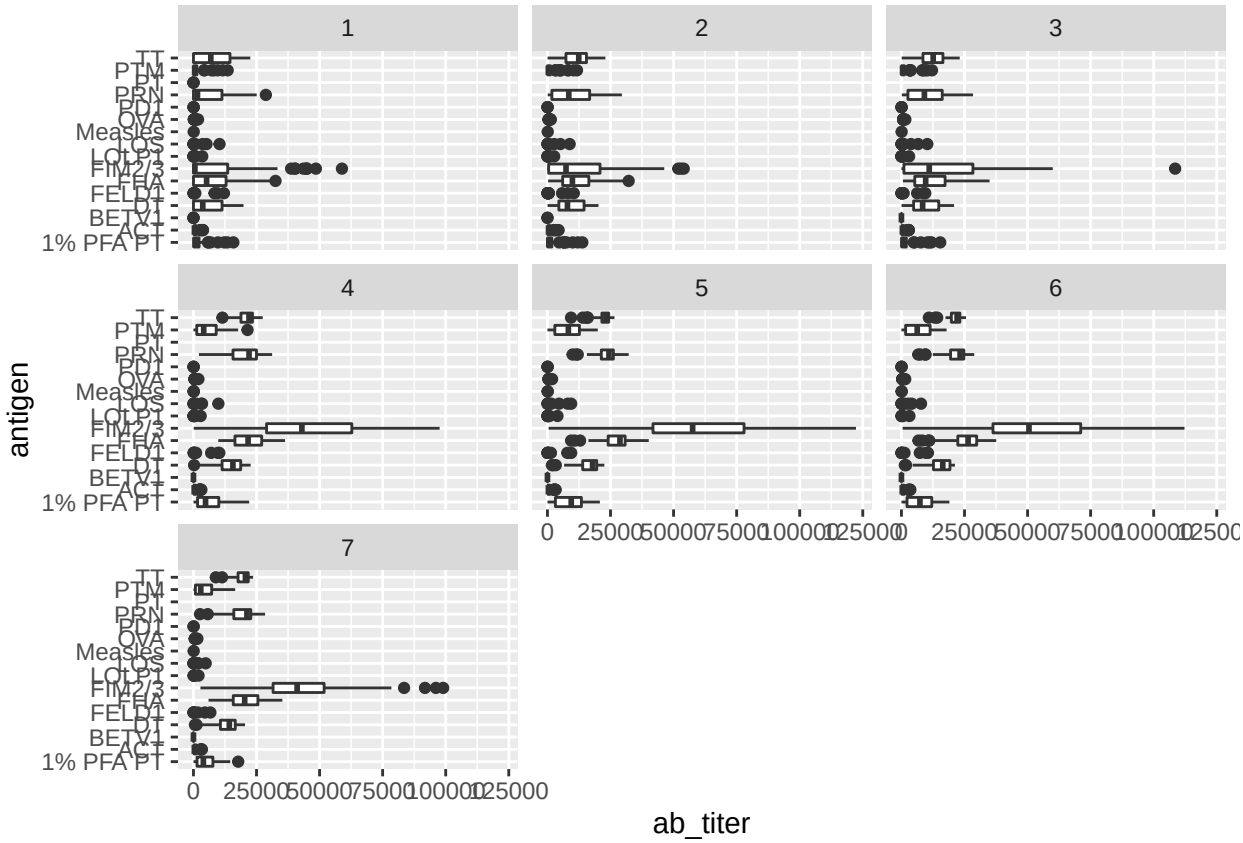
```
ig1 <- abdata %>% filter(isotype == "IgG1", visit!=8)
head(ig1)
```

```
##   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 infancy_vac biological_sex
## 1                 3.848750          1          wP         Female
## 2                 4.357917          1          wP         Female
## 3                 2.699944          1          wP         Female
## 4                 1.734784          1          wP         Female
## 5                 2.550606          1          wP         Female
## 6                 4.438966          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
```

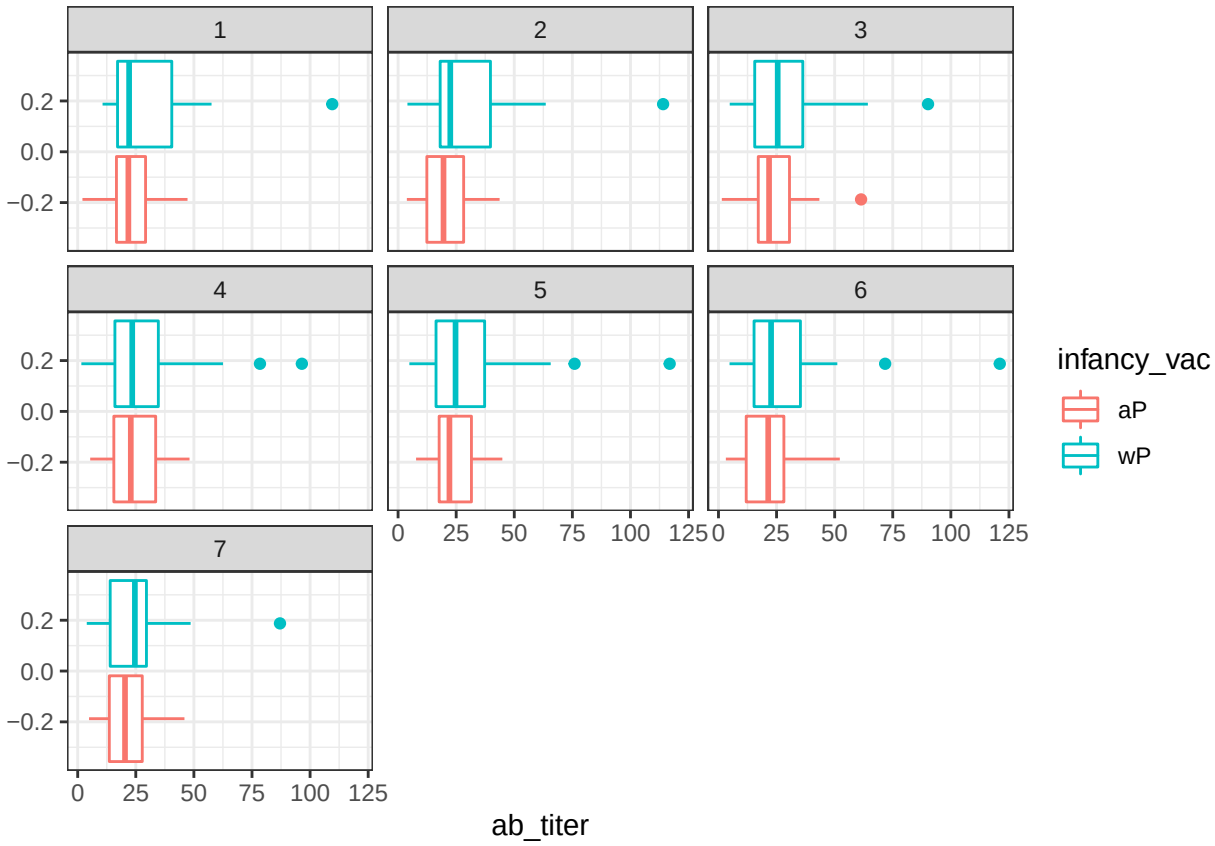
```
## 5 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
## 6 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
##   actual_day_relative_to_boost planned_day_relative_to_boost specimen_type
## 1                           -3                             0         Blood
## 2                           -3                             0         Blood
## 3                           -3                             0         Blood
## 4                           -3                             0         Blood
## 5                           -3                             0         Blood
## 6                           -3                             0         Blood
##   visit
## 1     1
## 2     1
## 3     1
## 4     1
## 5     1
## 6     1
```

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

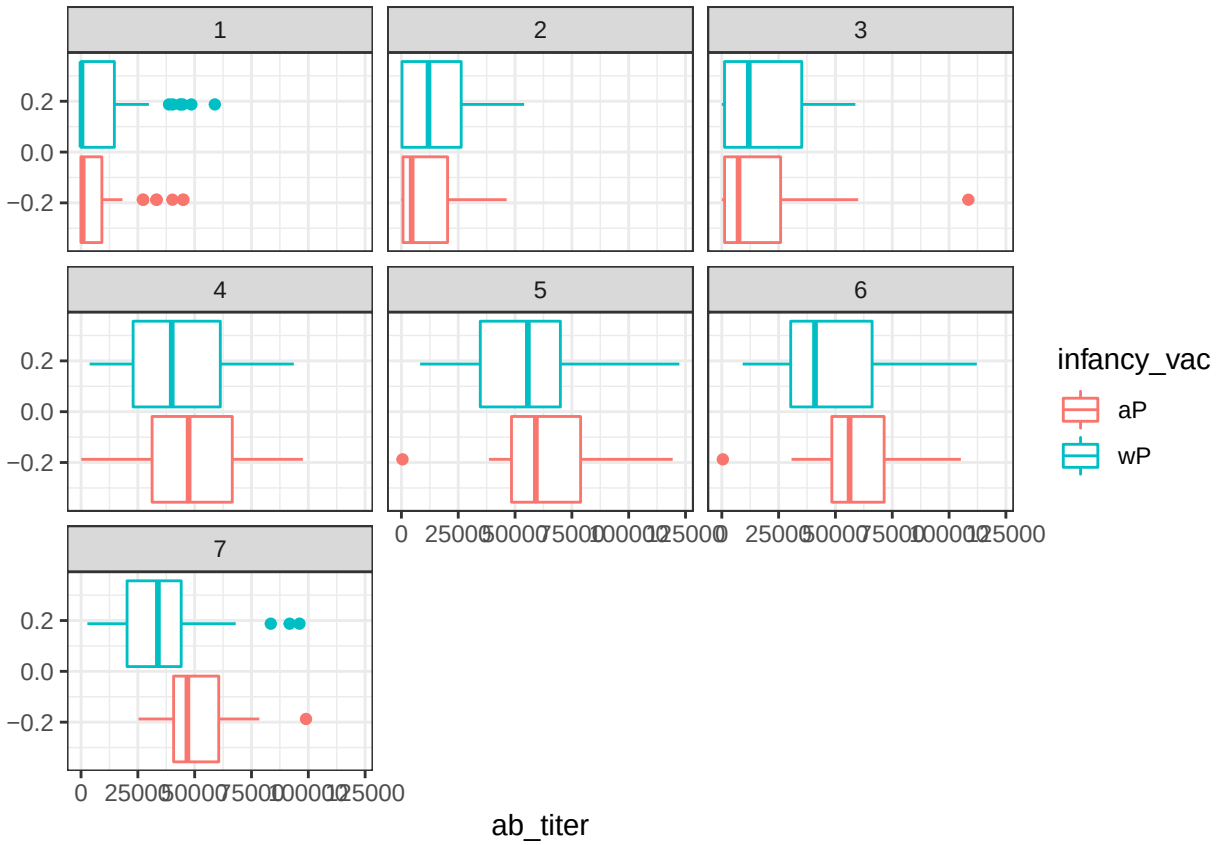
```
ggplot(ig1) +
  aes(ab_titer, antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow=3)
```



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




```
filter(ig1, antigen=="FIM2/3") %>%
  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 courses and the FIM2/3 data in particular?

Measles doesn't change much over time, while Fib2/3, as expected, increases dramatically over the visit time course.

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

Nothing very significant.

For RNA-Seq data the API query mechanism quickly hits the web browser interface limit for file size. We will present alternative download mechanisms for larger CMI-PB datasets in the next section. However, we can still do "targeted" RNA-Seq querys via the web accessible API.

For example we can obtain RNA-Seq results for a specific ENSEMBLE gene identifier or multiple identifiers combined with the & character:

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

rna <- read_json(url, simplifyVector = TRUE)
```

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

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

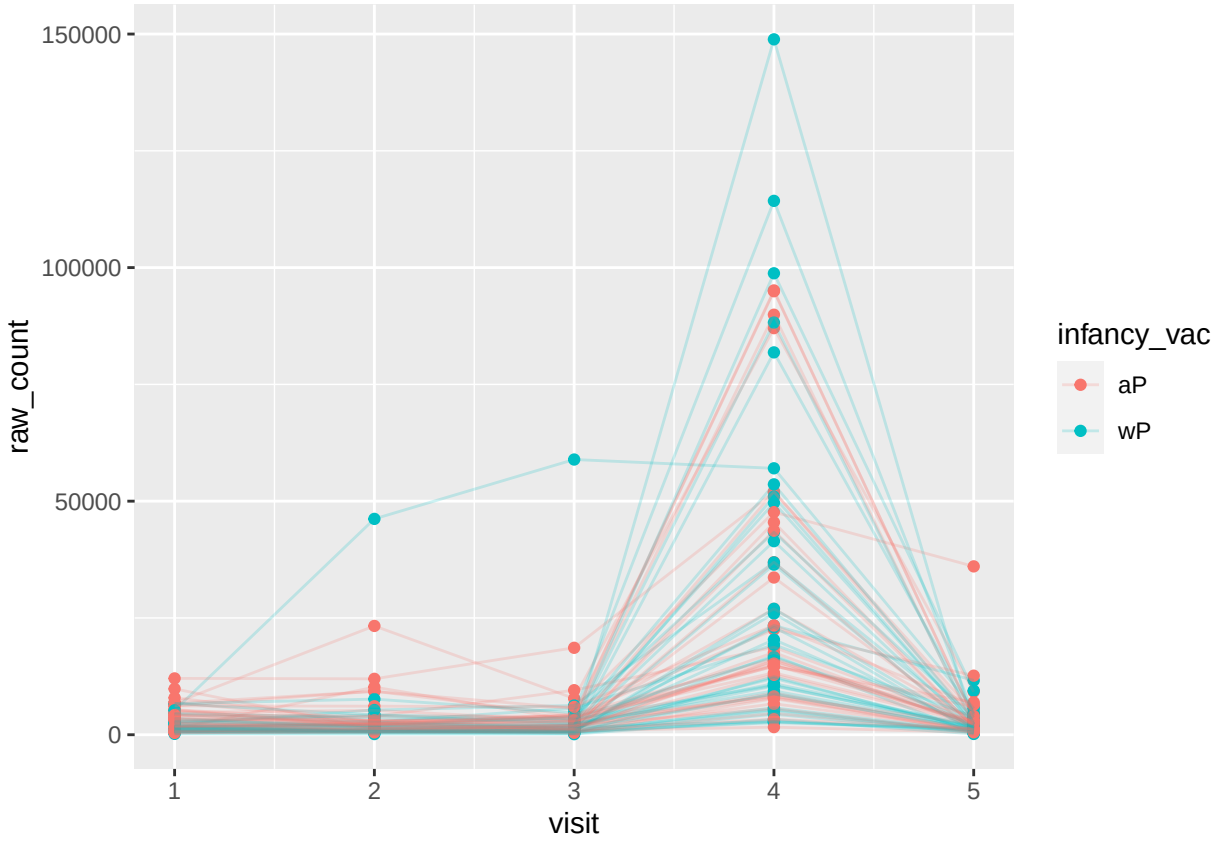
```
head(ssrna)
```

```
##   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
##   infancy_vac biological_sex              ethnicity               race
## 1          aP         Female     Hispanic or Latino More Than One Race
## 2          wP         Female Not Hispanic or Latino              Asian
## 3          wP           Male     Hispanic or Latino More Than One Race
## 4          aP         Female     Hispanic or Latino              White
## 5          aP         Female     Hispanic or Latino More Than One Race
## 6          wP         Female Not Hispanic or Latino              Asian
##   year_of_birth date_of_boost   study_name actual_day_relative_to_boost
## 1    1998-01-01    2016-11-07 2020_dataset                           3
## 2    1989-01-01    2016-09-26 2020_dataset                           3
## 3    1990-01-01    2016-10-10 2020_dataset                          15
## 4    1997-01-01    2016-10-24 2020_dataset                           1
## 5    1998-01-01    2016-11-07 2020_dataset                           7
## 6    1989-01-01    2016-09-26 2020_dataset                           7
##   planned_day_relative_to_boost specimen_type visit
## 1                             3         Blood     3
## 2                             3         Blood     3
## 3                            14         Blood     5
## 4                             1         Blood     2
## 5                             7         Blood     4
## 6                             7         Blood     4
```

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, raw_count, group=subject_id, color = infancy_vac) +
  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)?

UsuallyvVisit 4.

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

Not entirely. Antibody titer data peaks somewhat later, around visit 5, by which point gene expression level has already dropped. THis makes sense, as we'd expect a lag between transcription and translation, and we'd expect antibodies to persist, leading to a later peak.