Lab_16

RUNQI ZHANG

Table of contents

1. Investigating pertussis cases by year

```
# Import pakcages
library(datapasta) # allows directly copy&paste from website to R
```

```
Warning: package 'datapasta' was built under R version 4.2.2
```

```
library(ggplot2)
library(ggrepel)
library(jsonlite)  # read_json()
```

```
Warning: package 'jsonlite' was built under R version 4.2.2
```

```
library(lubridate) # working with date
```

```
Warning: package 'lubridate' was built under R version 4.2.2
```

```
Loading required package: timechange
```

```
Warning: package 'timechange' was built under R version 4.2.2
```

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

```
The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

```
library(dplyr)     # join()
```

```
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
```

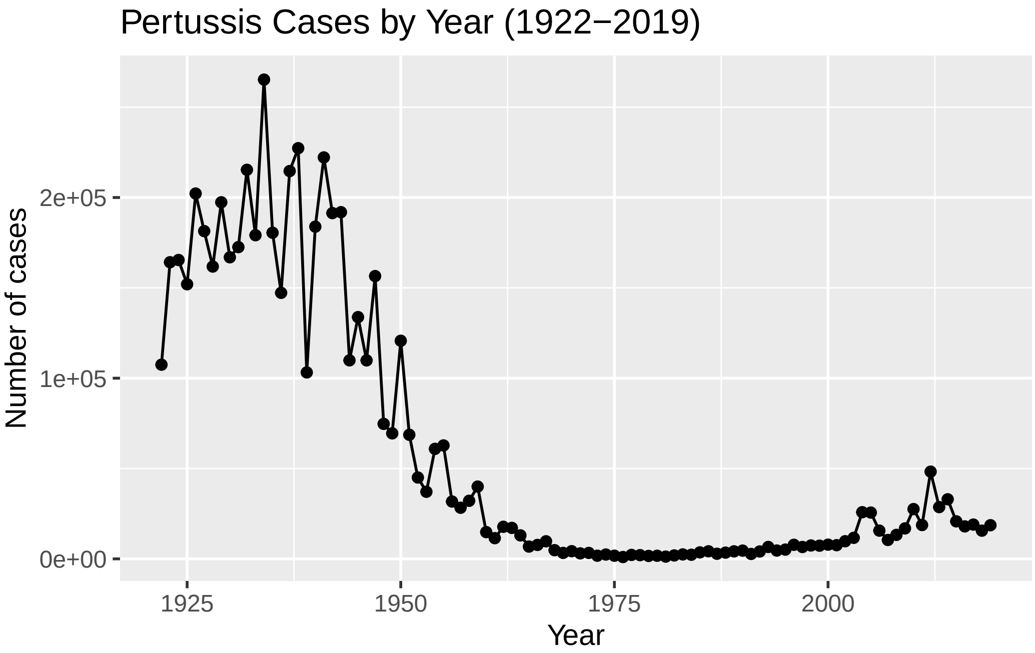
Q1. With the help of the R "addin" package datapasta assign the CDC pertussis case number data to a data frame called cdc and use ggplot to make a plot of cases numbers over time.

(link to datapasta manual: https://milesmcbain.github.io/datapasta/)

A: see plot below

```
# shortcuts for datapast(data.frame): ctrl + shift + D
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)
)

#rename second column
colnames(cdc)[2] = "case_number"
```

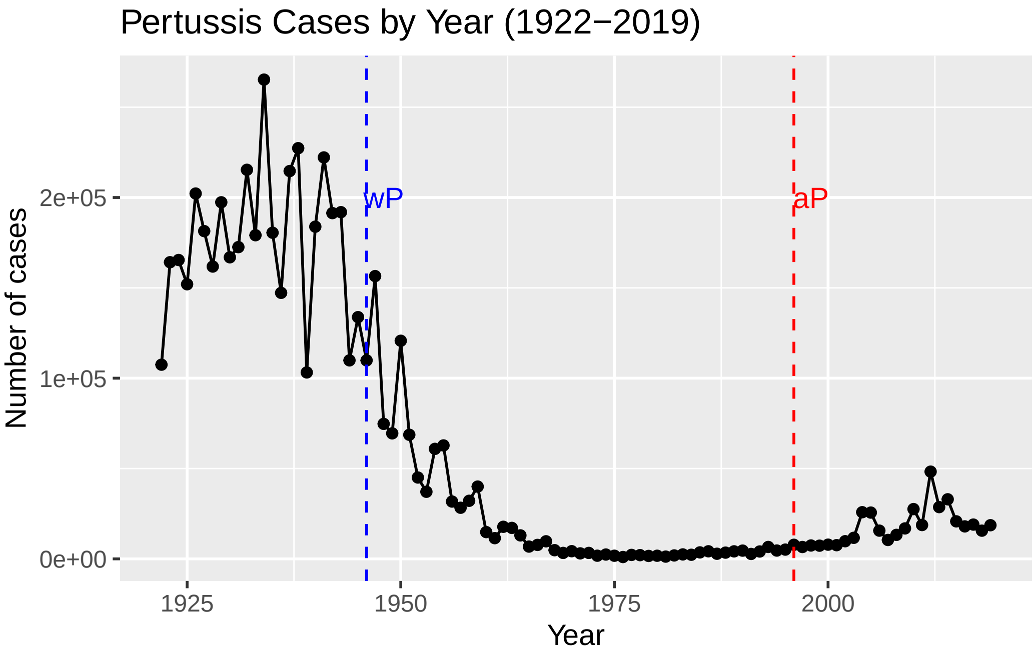
```
# ggplot of cases numbers over time
p1<-ggplot(cdc) +
      aes(Year, case_number) +
      geom_point() +
      geom_line() +
      labs(
        title="Pertussis Cases by Year (1922-2019)",
        x="Year",
        y="Number of cases"
      )
p1
```



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

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?

```
p1+geom_vline(xintercept=1946, col="BLUE", linetype="dashed")+annotate("text", label="wP",
   geom_vline(xintercept=1996, col="RED", linetype="dashed")+annotate("text", label="aP",
```



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

A: According to the cdc data, puertussis cases showed resurgence after the introduction of aP vaccination. Possible explanations are 1)application of more sensitive testing methods, 2)growing vaccination hesitancy, 3)bacterial evolution leading to inefficacy of vaccine, and 4)short-lived immunity of aP vaccine as compared to wP.

3. Exploring CMI-PB data

```
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      dataset
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?

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

```
aP wP
47 49
```

A: aP-47 wP-49

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

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

```
Female   Male
    66     30
```

A: Female-66 Male-30

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

```r
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
```

A: see table above

Investigating age effect

```
# accessing age info
age <- time_length( today()-ymd(subject$year_of_birth), "years")
age <- round(age, 0) # round age to interger

wP_subject <- subject[subject$infancy_vac=="wP",]
wP_age <- time_length( today()-ymd(wP_subject$year_of_birth), "years")
wP_age <- round(wP_age, 0) # round age to interger

aP_subject <- subject[subject$infancy_vac=="aP",]
aP_age <- time_length( today()-ymd(aP_subject$year_of_birth), "years")
aP_age <- round(aP_age, 0) # round age to interger
```

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?

```
mean(wP_age)
```

```
[1] 36.16327
```

```
mean(aP_age)
```

```
[1] 25.31915
```

```
t.test(wP_age,aP_age)
```

```

	Welch Two Sample t-test

data:  wP_age and aP_age
t = 12.092, df = 51.084, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  9.043752 12.644481
sample estimates:
mean of x mean of y 
 36.16327  25.31915 
```

A: i)the average age of wP individuals is 36; ii)the average age of aP individuals is 25; iii)their ages are significantly different according to welch 2-sample t-test (p-value < 2.2e-16).

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

```
age_boost <- time_length( ymd(subject$date_of_boost)-ymd(subject$year_of_birth), "years")
age_boost <- round(age, 0) # round age to interger
mean(age_boost)
```

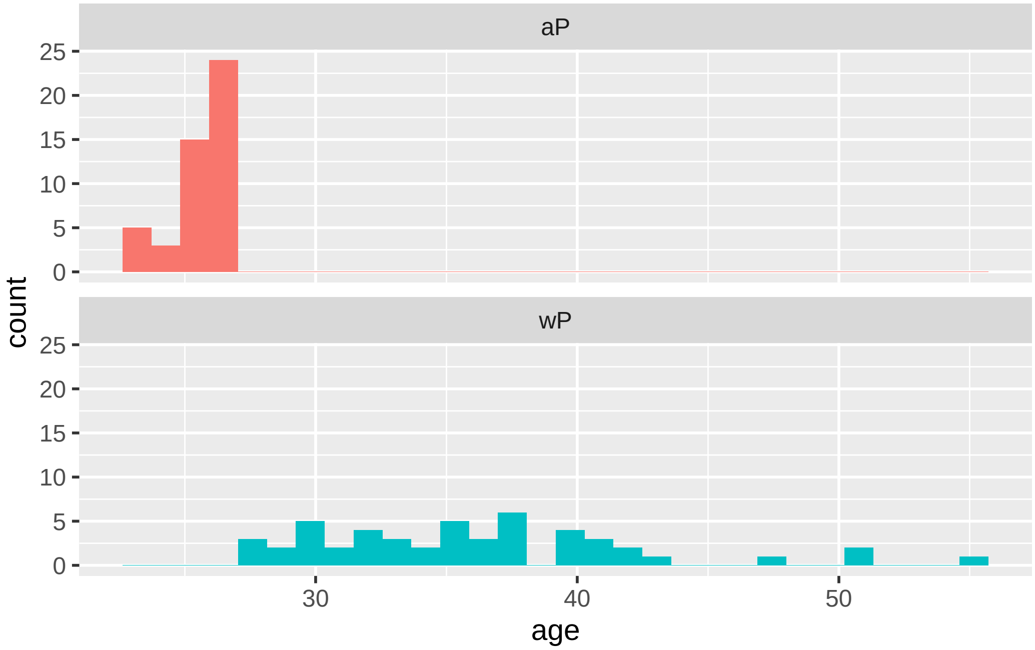
```
[1] 30.85417
```

A: the age of all individuals at time of boost are calculated and saved under variable_name "age_boost"; the average age of receiving boost is 31 (30.85).

Q9. With the help of a faceted boxplot (see below), do you think these two groups are significantly different?

```
ggplot(subject) +
  aes(age,
      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`.
```



A: Based on the plot, I think the two groups are significantly different in ages

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)
```

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      dataset
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    20
```

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 
```

A: the number of visit for specimen #8 is significantly fewer than other specimens

4. 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.

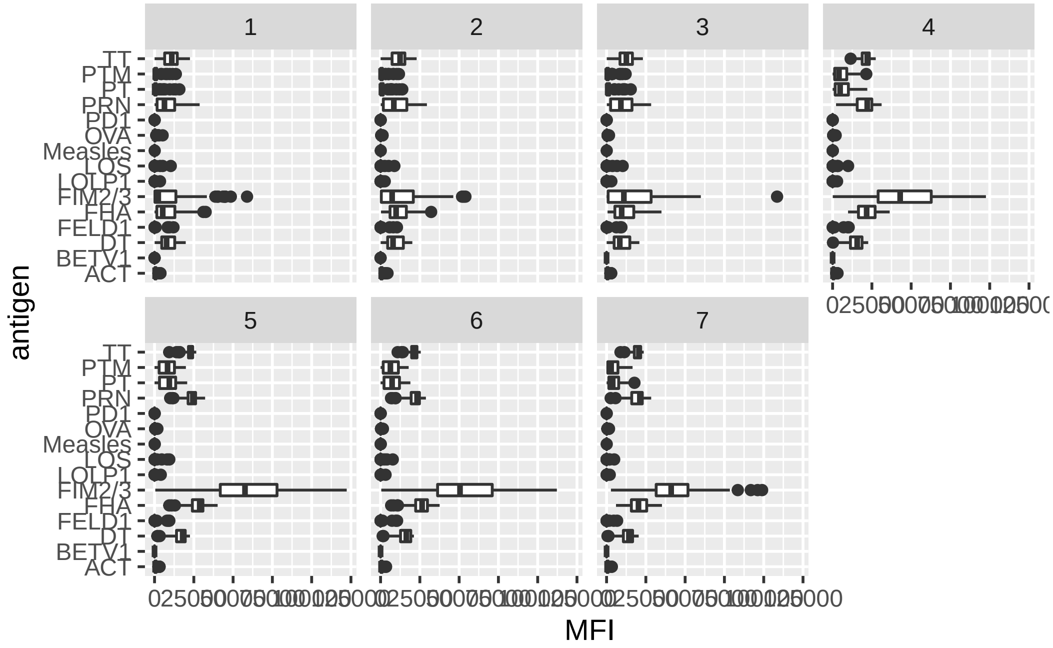
```
ig1 <- abdata %>% filter(isotype == "IgG1", visit!=8)
head(ig1)
```

```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1    IgG1                TRUE     ACT 274.355068      0.6928058
2           1    IgG1                TRUE     LOS  10.974026      2.1645083
3           1    IgG1                TRUE   FELD1   1.448796      0.8080941
4           1    IgG1                TRUE   BETV1   0.100000      1.0000000
5           1    IgG1                TRUE   LOLP1   0.100000      1.0000000
6           1    IgG1                TRUE Measles  36.277417      1.6638332
   unit lower_limit_of_detection subject_id actual_day_relative_to_boost
1 IU/ML                 3.848750          1                           -3
2 IU/ML                 4.357917          1                           -3
3 IU/ML                 2.699944          1                           -3
4 IU/ML                 1.734784          1                           -3
```

```
5 IU/ML                 2.550606          1                           -3
6 IU/ML                 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      dataset
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(MFI, antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow=2)
```

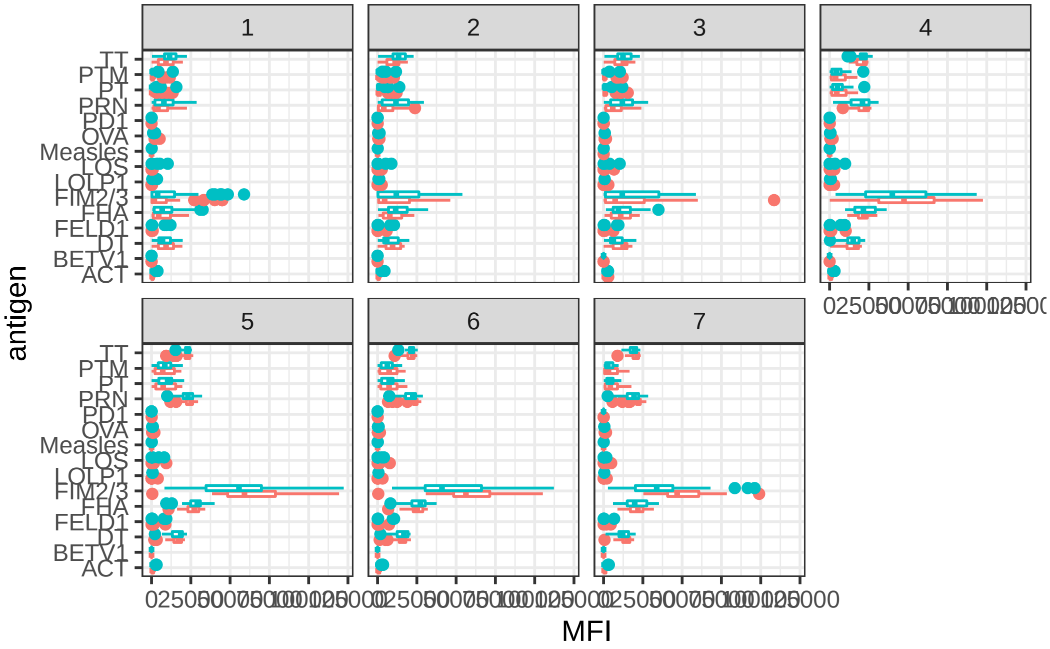


Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

A: FIM2/3 antigens show differences in the level of IgG1 antibody titer, indicating that this antigen is involved in the pathway and worthes further investigation.

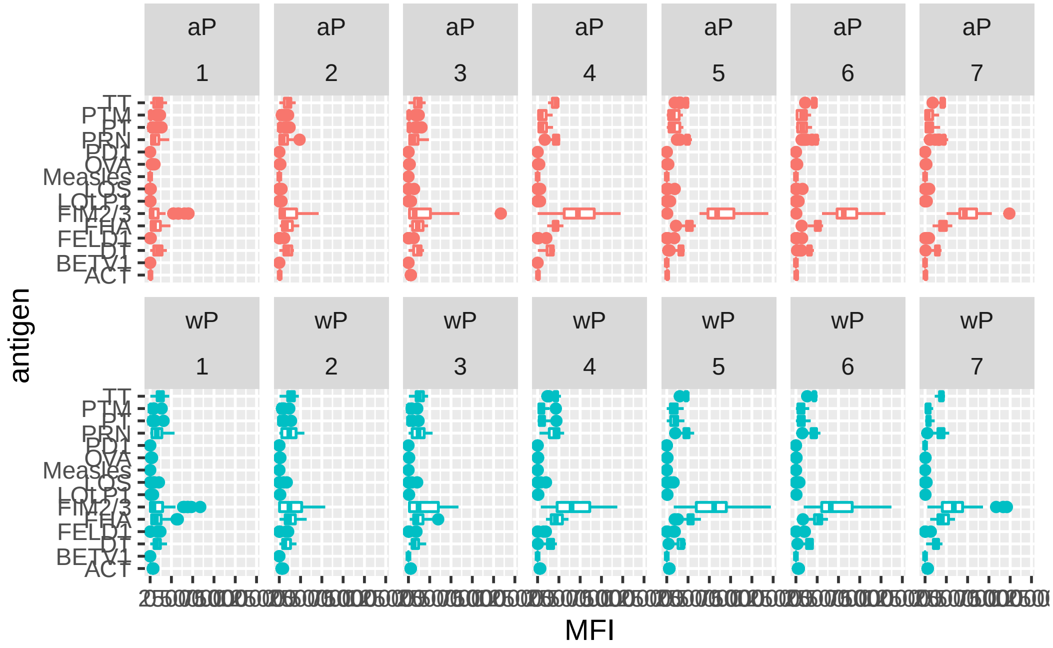
We can attempt to examine differences between wP and aP here by setting color and/or facet values of the plot to include infancy_vac status (see below). However these plots tend to be rather busy and thus hard to interpret easily.

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



Another version of this plot adding infancy_vac to the faceting:

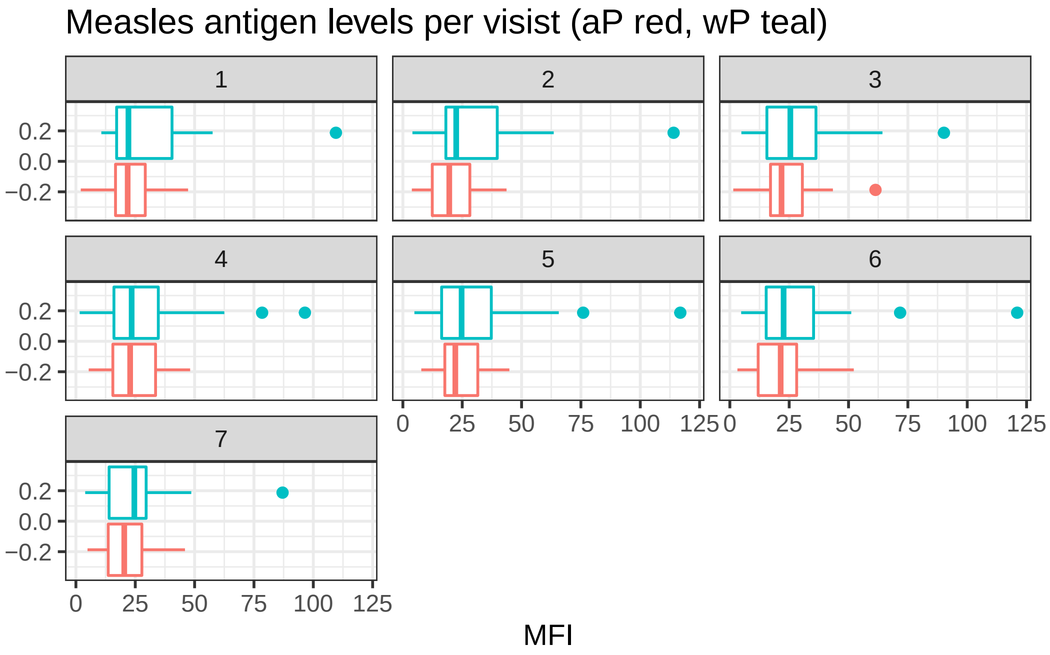
```
ggplot(ig1) +
  aes(MFI, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```



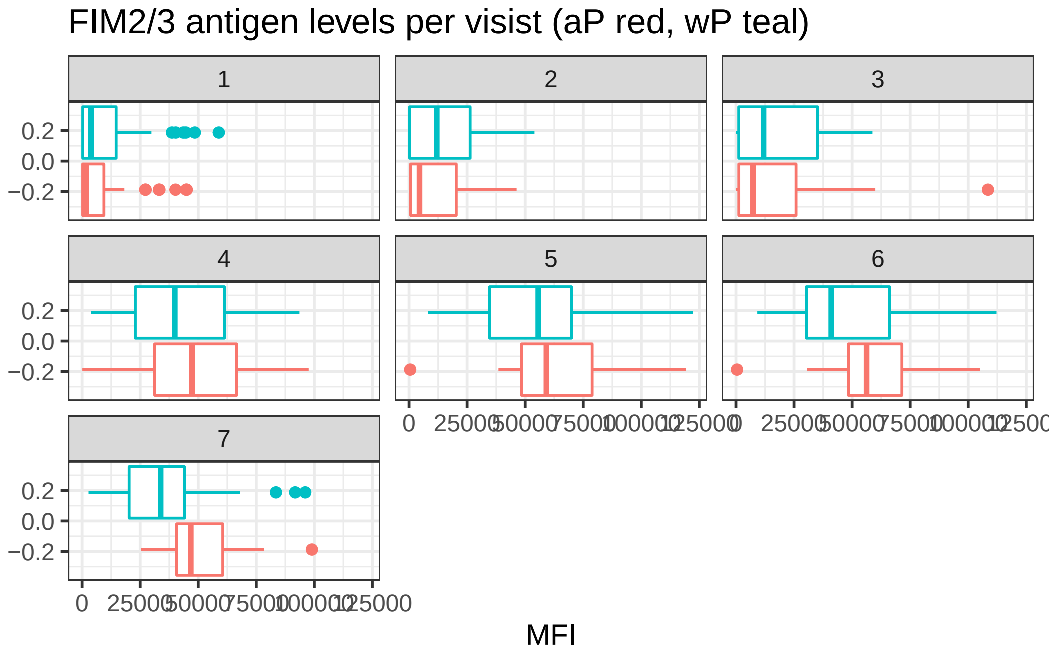
Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each.

control antigen: Measles (not in our vaccines); antigen of interest: FIM2/3 (extra-cellular fimbriae proteins from B. pertussis that participate in substrate attachment).

```
filter(ig1, antigen=="Measles") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw()+
  labs(
    title="Measles antigen levels per visist (aP red, wP teal)"
  )
```



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



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

A: Measles antigen levels remained unchanged over time and are consistent across the two groups; FIM2/3 antigen levels increased over time, the rate of increase is consistent across groups,and its level remained high for a longer period of time in the aP subjects.

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

A: A clear difference in aP vs wP responses can be found at the FIM2/3 antigen level during the 7th visit. At the time of the visit, antigen level began to show a tendency to decrease in wP subjects, while its level remained high in aP subjects.

5. Obtaining CMI-PB RNASeq data

For example use the following URL (https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.E

The link above is for the key gene involved in expressing any IgG1 antibody, namely the IGHG1 gene. Let's read available RNA-Seq data for this gene into R and investigate the time course of it's gene expression values.

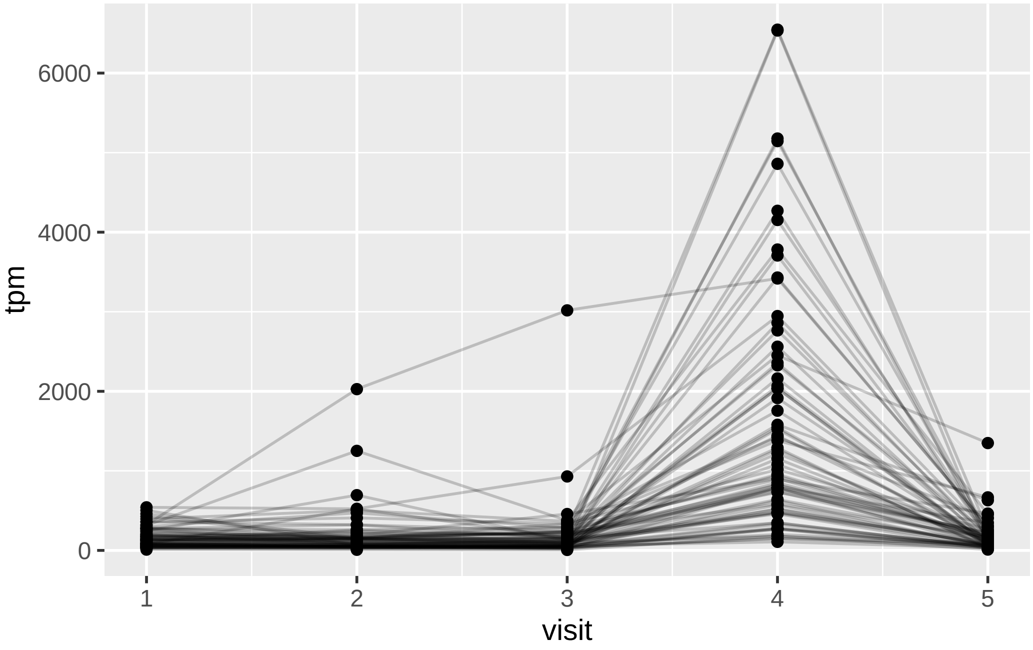
```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.

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

```
#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)?

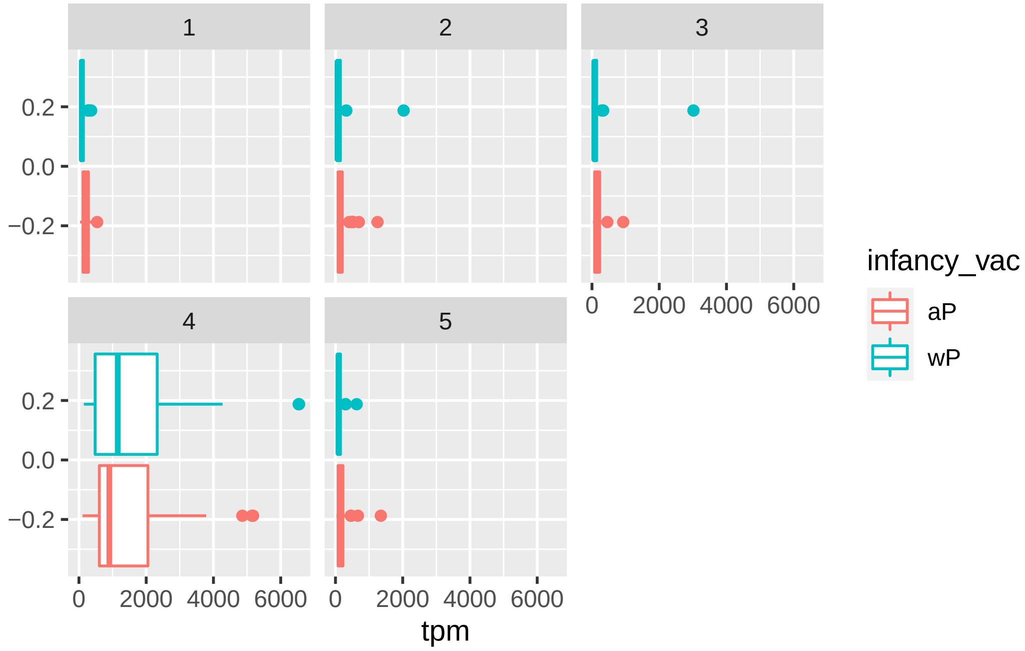
A: the peak gene expression is found during the 4th visit, and the expression levels show variations among subjects. The expression level in the majority of subjects remained low till the 3rd visit, and drop back to its previously seen low level at the 5th visit.

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

A: the rising phase of the expression pattern match the trend of antibody titer data, however, their falling phase disagrees, indicating the reduction in FIM2/3 antigen levels is regulated by other gene(s).

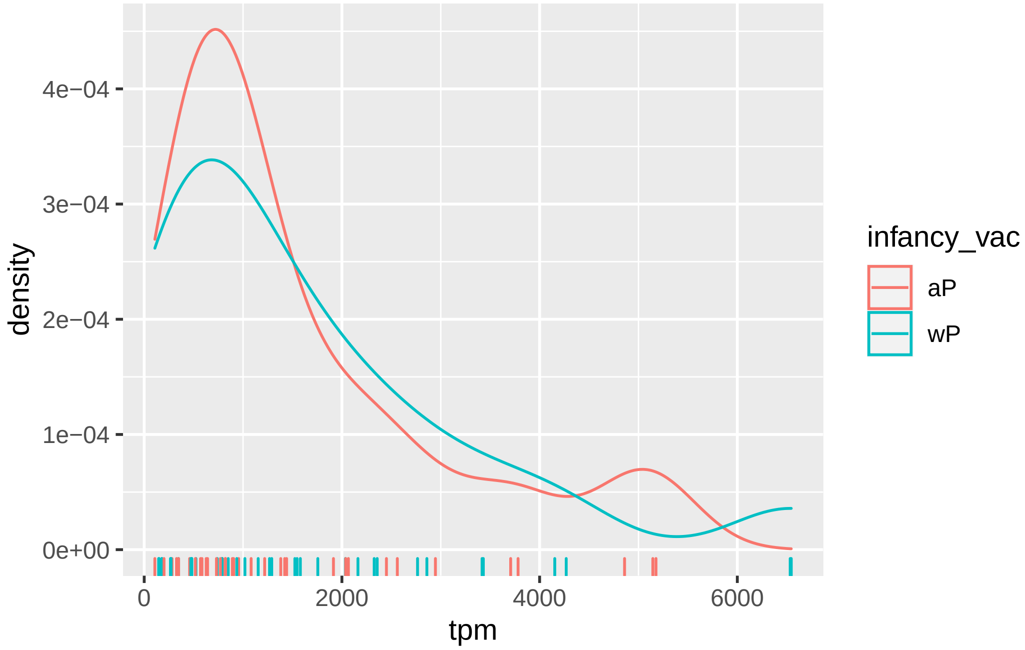
We can dig deeper and color and/or facet by infancy_vac status:

```
ggplot(ssrna) +
  aes(tpm, col=infancy_vac) +
  geom_boxplot() +
  facet_wrap(vars(visit))
```



There is no obvious wP vs. aP differences here even if we focus in on a particular visit:

```
ssrna %>%
  filter(visit==4) %>%
  ggplot() +
    aes(tpm, col=infancy_vac) + geom_density() +
    geom_rug()
```


density
4e−04
3e−04
2e−04
1e−04
0e+00
tpm
infancy_vac
aP
wP