Class19

Gregory Jordan

1. Investigating pertussis cases by 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.

Scrape the CDC data on pertusis cases per year in the US from their website here: https://www.cdc.gov/pertussis/surv-reporting/cases-by-year.html

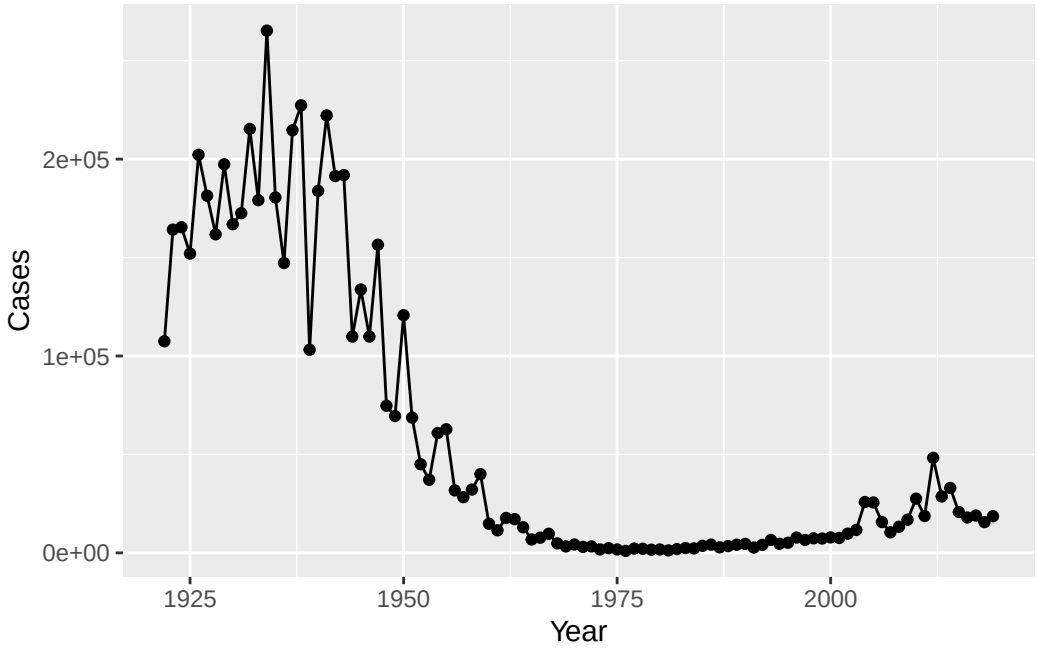
```
library(datapasta)

cdc <- data.frame(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),
                    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(data=cdc) + aes(x=Year,y=Cases) + geom_line() + geom_point()
```

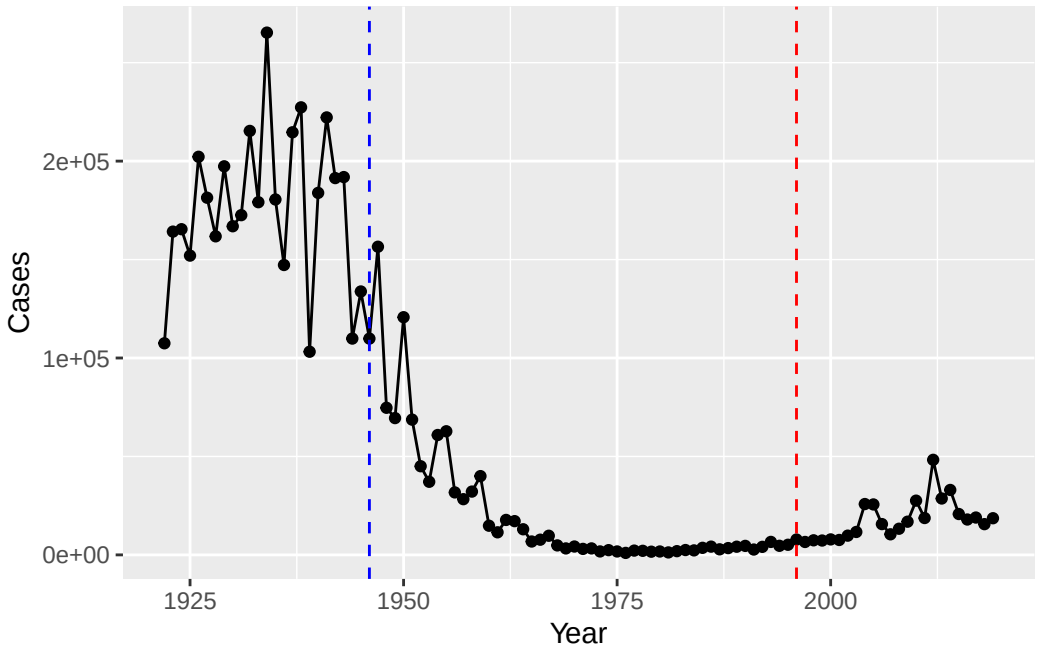


2. A tale of 2 vaccines

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?

the wP vaccine gets introduced and then the cases start to plummet towards 0 and stay at 0 until they surge up again after the next vaccine

```
ggplot(data=cdc) + aes(x=Year,y=Cases) + geom_line() + geom_point() + geom_vline(xintercep
geom_vline(xintercept=1996,col="red",linetype="dashed")
```



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

The amount of cases increased. I think a big factor was the rise of antivax movement and people not trusting the new vaccine. People didn't trust it and then ended up not getting it and more likely to get sick. Also, people probably thought they did not need a new vaccine because cases were so low for a long time and so people decided getting a new vaccine was uneccesary.

The pertussis field has several hypotheses for the resurgence of pertussis including (in no particular order): 1) more sensitive PCR-based testing, 2) vaccination hesitancy 3) bacterial evolution (escape from vaccine immunity), 4) waning of immunity in adolescents originally primed as infants with the newer aP vaccine as compared to the older wP vaccine.

3. Exploring CMI-PB data

Json Data

The CMI-PB API returns JSON data. need to read the JSON key-value pairs using the `read_json()` function. simplifies key-value pairs into R dataframes

```
library(jsonlite)
subject <- read_json("https://www.cmi-pb.org/api/subject", simplifyVector = TRUE)
#simplifyvector argument makes it so you get a table. if it was FALSE you would be getting
```

```
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 many aP and wP infancy vaccinated subjects are in the dataset? Solution

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

```
aP wP
47 49
```

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$ethnicity,subject$biological_sex)
```

```
                        Female Male
  Hispanic or Latino        18    5
  Not Hispanic or Latino    47   22
  Unknown                    1    3
```

Working with Dates in the Data Frame:

Working with Dates

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?

(i)

```
library(lubridate)
```

```
Loading required package: timechange
```

```

Attaching package: 'lubridate'
```

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

    date, intersect, setdiff, union
```

```
time_length(today()-mean(ymd(subject[subject$infancy_vac=="wP",]$year_of_birth)),"years")
```

```
[1] 36.07532
```

(ii)

```
time_length(today()-mean(ymd(subject[subject$infancy_vac=="aP",]$year_of_birth)),"years")
```

```
[1] 25.23087
```

(iii)

```
cat(36-25,"years difference between mean age of wP and aP individuals")
```

```
11 years difference between mean age of wP and aP individuals
```

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

```
#adding in subject age
subject$age <- time_length(today()-ymd(subject$year_of_birth),"years")
```

```
subject$age_at_boost <- time_length((today()-ymd(subject$date_of_boost)),"years")
```

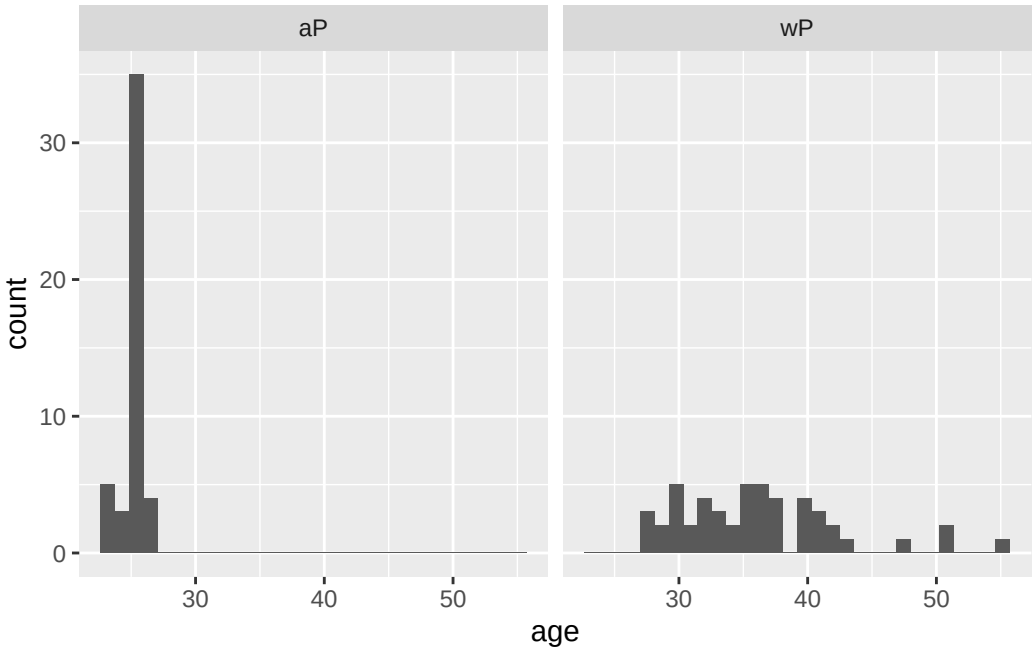
```
head(subject)
```

```
  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      dataset      age age_at_boost
1    1986-01-01    2016-09-12 2020_dataset 36.91170     6.214921
2    1968-01-01    2019-01-28 2020_dataset 54.91307     3.838467
3    1983-01-01    2016-10-10 2020_dataset 39.91239     6.138261
4    1988-01-01    2016-08-29 2020_dataset 34.91307     6.253251
5    1991-01-01    2016-08-29 2020_dataset 31.91239     6.253251
6    1988-01-01    2016-10-10 2020_dataset 34.91307     6.138261
```

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

```
ggplot(subject) + aes(age) + geom_histogram() + facet_wrap(vars(infancy_vac))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Yes. These groups are significantly different b/c the aP vax is given way earlier in age than the wP vax.

Joining multiple tables

read specimen and ab_titer tanbles into R and store them as specimen and titer

```
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:

we want to join

```
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
```

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

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

```
dim(meta)
```

```
[1] 729  15
```

```
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     age
1 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.9117
2 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.9117
3 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.9117
4 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.9117
5 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.9117
6 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.9117
```

```
  age_at_boost
1     6.214921
2     6.214921
3     6.214921
4     6.214921
5     6.214921
6     6.214921
```

```
#I have an extra column because I added age of subject when they got the boost column
```

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    22
```

```
head(abdata)
```

```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1     IgE               FALSE   Total 1110.21154       2.493425
2           1     IgE               FALSE   Total 2708.91616       2.493425
3           1     IgG                TRUE      PT   68.56614       3.736992
4           1     IgG                TRUE     PRN  332.12718       2.602350
5           1     IgG                TRUE     FHA 1887.12263      34.050956
6           1     IgE                TRUE     ACT    0.10000       1.000000
   unit lower_limit_of_detection subject_id actual_day_relative_to_boost
1 UG/ML                 2.096133          1                           -3
2 IU/ML                29.170000          1                           -3
3 IU/ML                 0.530000          1                           -3
4 IU/ML                 6.205949          1                           -3
5 IU/ML                 4.679535          1                           -3
6 IU/ML                 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      dataset     age
1 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.9117
2 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.9117
3 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.9117
4 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.9117
5 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.9117
6 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset 36.9117
  age_at_boost
1     6.214921
2     6.214921
3     6.214921
4     6.214921
5     6.214921
6     6.214921
```

```
#I have extra column because I added some age columns to my subject data frame
```

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
```

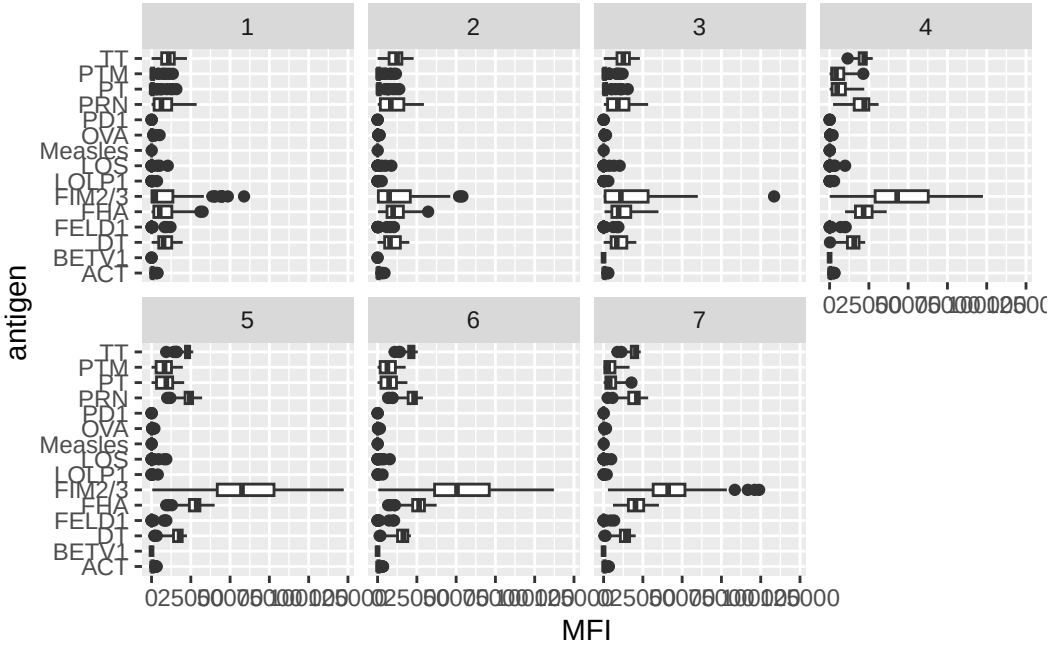
way fewer visit 8 specimens!!

4. Examine IgG1 Ab titer levels

```
#filter for igG1 and exclude small visit group
ig1 <- abdata %>% filter(isotype == "IgG1", visit != 8)
```

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

```
#plotting MFI by antigen and facet wrap to group by visit time
ggplot(ig1) +
  aes(MFI, antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow=2)
```



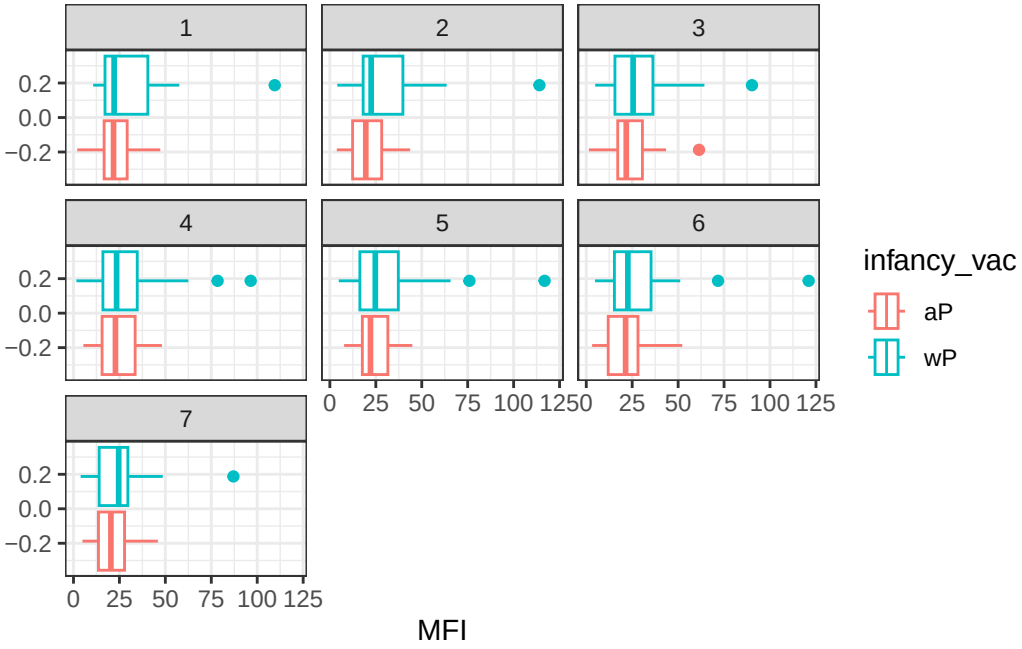
Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

FIM 2/3 is the antigen with the strongest increase over time. Also, TT, PRN, DT, FHA, and a couple others. Things like Measles, OVA, PD1, LOS, LOLP1, etc. do not increase at all. This makes sense as the antigens that are increasing are antigens associated with pertussis/dTAP vaccine whereas the antigens not increasing are not associated with it (like measles, etc.) It

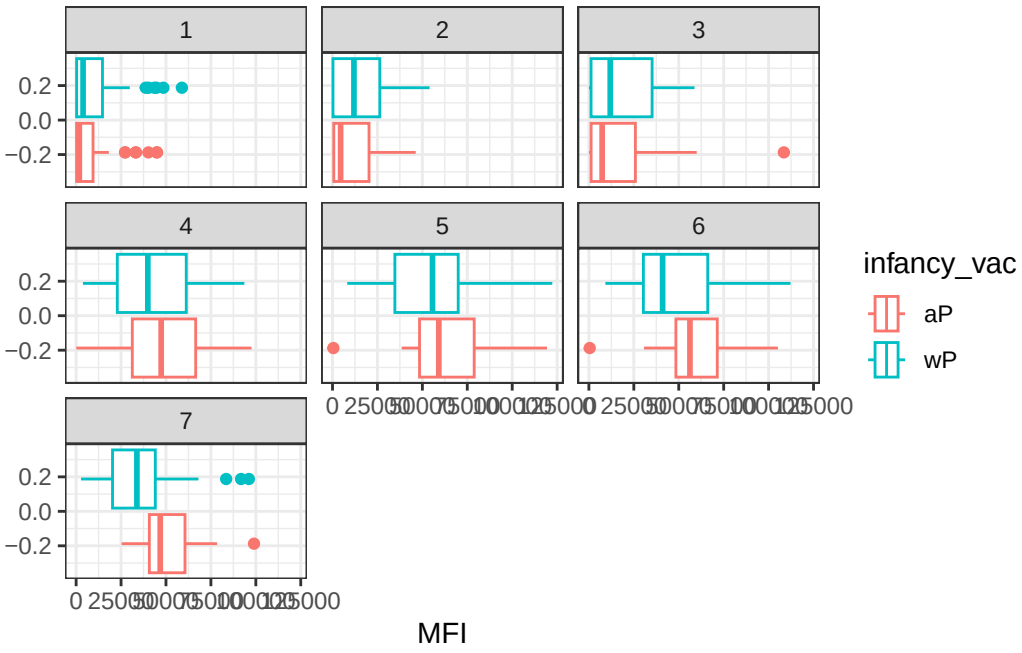
is cool to see that pertussis antigens/dTAP recognized by IgG1 antibodies are increasing over time. This is showing the antibody response to the vaccine (or an infection)!

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(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = TRUE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```




```
filter(ig1, antigen=="FIM2/3") %>%
  ggplot() +
  aes(MFI, 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 stays the same over tie whereas FIM2/3 MFI increases over time! They peak around visit 5 and then start to drop. This seems to be a sign that the IgG1 response surges up then hits a peak around time point 5 and then starts to drop off, which would make sense for IgG1 response

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

aP peak around visit 5 gets higher than wP peak (meaning higher IgG1 response to vaccine) and the IgG1 titer lasts longer in aP whereas the wP starts to drop off faster around day 6/7

5. Obtaining CMI-PB RNA-Seq Data

```
#for IGHG1 gene
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.

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

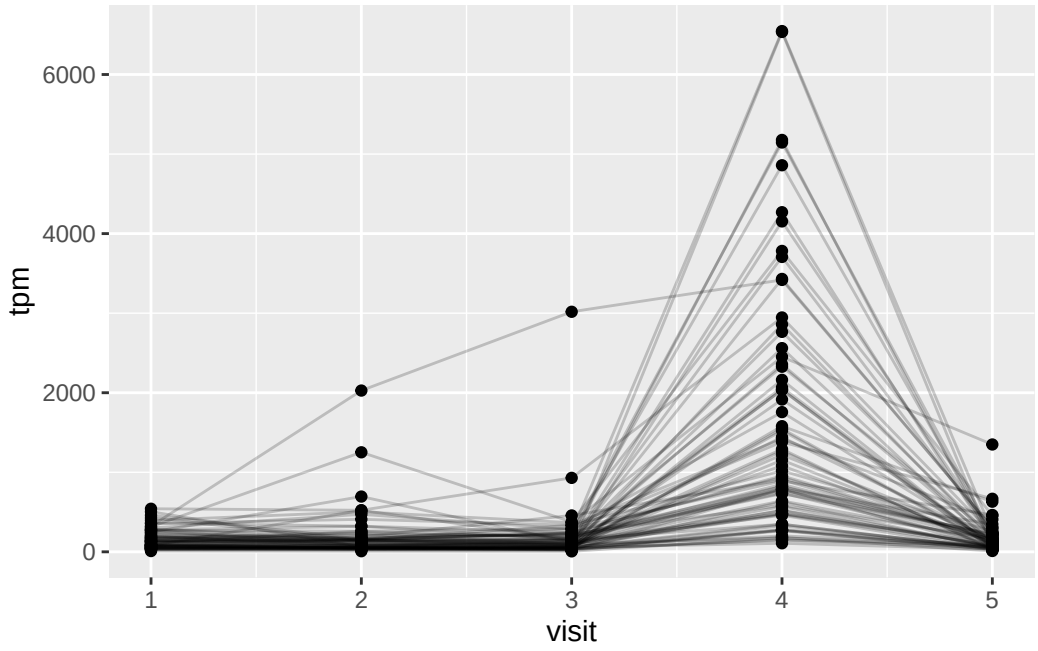
```
#join RNA-seq data with metadata

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(data=ssrna) + aes(x=visit,y=tpm,group=subject_id) + geom_point() + geom_line(alpha=
```



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

at max level at visit 4. it typically spikes really quickly at visit 4 and then drops down by visit 5

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

This does follow the same trend of surging up towards visit 4. This makes sense as the IGHG1 gene (encoding for immunoglobulin heavy chain gamma 1) would be activated to make antibodies. So we see this gene transcription getting upregulated basically right before

the antibodies are being made, and then the gene promptly shuts off, and the antibodies stay in the system/slowly decrease in titer, but much slower because they are long lived. This is cool! We're seeing a gene turning on and off and seeing an increase in the antibodies it is making all during a time course.