Pertussis_MiniProject

Jordan Prych (A17080226)

Table of contents

Pertussis (aka whooping cough) is a deadly respiratory lung infection caused by the bacteria B. Pertussis.

The CDC tracks Pertussis cases around the US. https://www.cdc.gov/pertussis/surv-reporting/cases-by-year.html

We can "scrape" this data using the R **datapasta** package.

```r
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,2020L,2021L,2022L,2024),
 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,
                                   6124,2116,3044,35493)
)
```

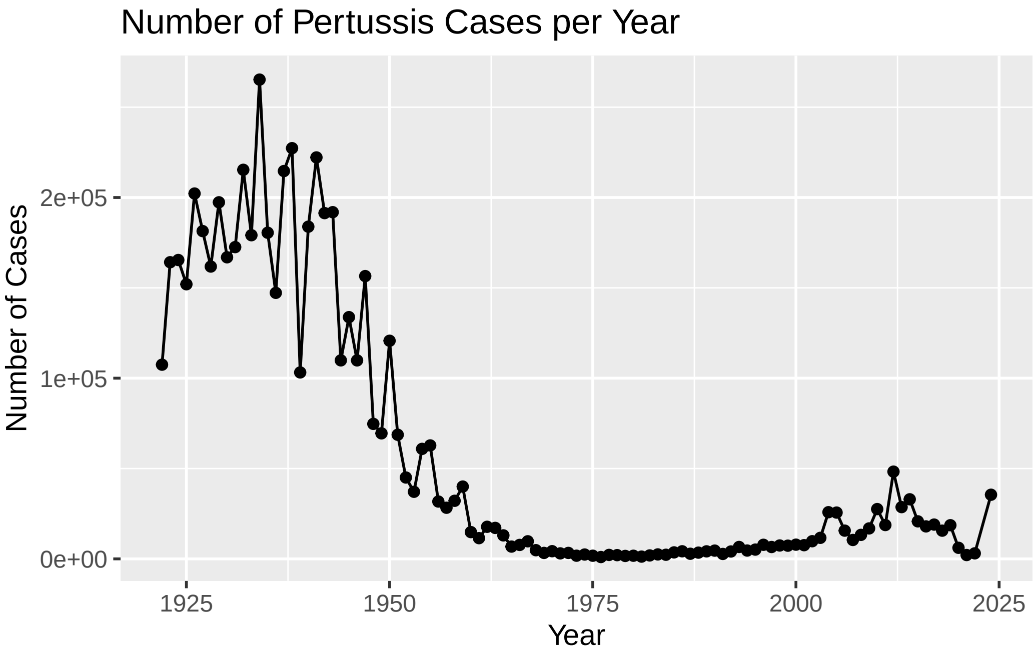
```
head(cdc)
```

```
  year  cases
1 1922 107473
2 1923 164191
3 1924 165418
4 1925 152003
5 1926 202210
6 1927 181411
```

> 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.

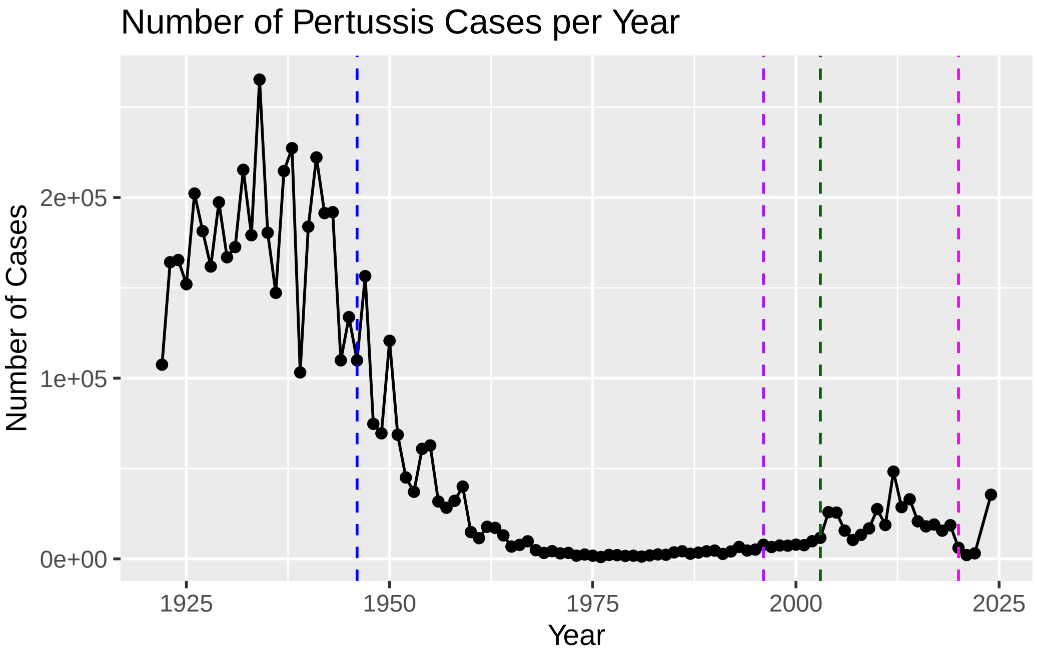
```
library(ggplot2)

ggplot(cdc) + aes(year, cases) + geom_point() + geom_line() + xlab("Year") + ylab("Number of
```



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, cases) + geom_point() + geom_line() + xlab("Year") + ylab("Number of
```



There were high case numbers before the first wP (whole-cell) vaccine in 1946(blue line). Then there was a rapid decline in case numbers until 2004(green line) when we have our first large-scale outbreaks of pertussis again. There is also a noticeable COVID-related dip and reacent rapid rise.

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

After the introduction of the aP vaccine in 1996(purple line), there were low case numbers until a rise in 2004. There are many possible explanations for this occurance inclusing the idea that the aP vaccine causes wavering immunity leading to a spike in cases years later, causing the requirement for a booster shot in comaprison to the older wP vaccine.

Big Question: what is different about the immune response to infection if you have an older wP vaccine versus the newer aP vaccine? Is it the vaccine's fault?

There is no definite answer to this question yet.

Exploring CMI-PB Data

CMI- Computational Models of Immunity- Pertussis Boost

The CMI-PB project aims to address this key question: what is the difference between aP and wP individuals.

We can get all the data from this ongoing project via JSON API calls. For this we will use the **jsonlite** package. We can install with `install.packages("jsonlite")`

```r
library(jsonlite)
```

```
Warning: package 'jsonlite' was built under R version 4.4.3
```

```r
subject <- read_json("https://www.cmi-pb.org/api/v5_1/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      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
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
```

> Q How many indiividuals "subjects" are in this dataset?

```r
nrow(subject)
```

```
[1] 172
```

> Q4. How many wP and aP primmed individuals are in this dataset?

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

```

aP wP 
87 85 
```

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

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

```
Female   Male 
   112     60 
```

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

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

```
                                           
                                            Female Male
  American Indian/Alaska Native                  0    1
  Asian                                         32   12
  Black or African American                      2    3
  More Than One Race                            15    4
  Native Hawaiian or Other Pacific Islander      1    1
  Unknown or Not Reported                       14    7
  White                                         48   32
```

Side-Note: Working with Dates

Two columns of `subject` contain dates in Year-Month-Day format. Using the **lubricate** package we can eaily work with dates in this format.

```r
library(lubridate)
```

```

Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

What is today's date?

```
today()
```

```
[1] "2025-03-09"
```

How many days have passes since new year 2000?

```
today()-ymd("2000-01-01")
```

```
Time difference of 9199 days
```

What is this in years?

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

```
[1] 25.18549
```

use `ymd()` function to tell lubricate the format of our particular date and then use `time_legnth()` function to convert days to years

> 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?

```
subject$age <- today()- ymd(subject$year_of_birth)
```

```
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
```

```r
# average age of aP individuals
ap <- subject %>% filter(infancy_vac == "aP")

round( summary( time_length( ap$age, "years" ) ) )
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     22      26      27      27      28      34
```

```r
#average age of wP individuals

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

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     22      32      34      36      39      57
```

These results are not significantly different because the minimum and maxmimum values(the range) for ap and wp are not different, meaning that they overlap. Since the ranges overlap, the average age is not significantly different.

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

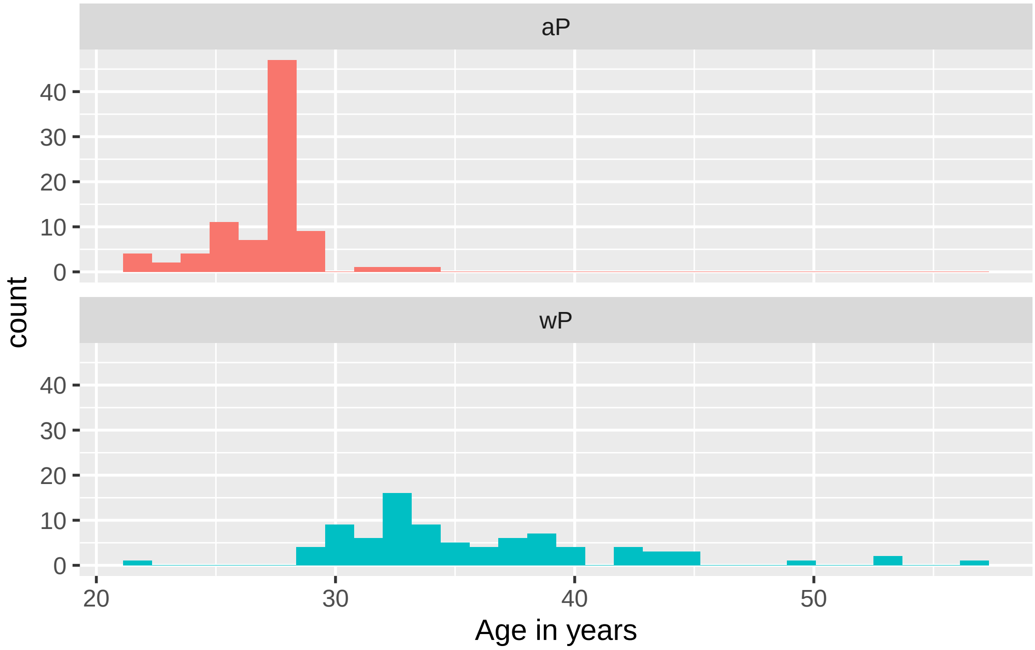
```r
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

```
[1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481
```

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

```r
ggplot(subject) +
  aes(time_length(age, "year"),
      fill=as.factor(infancy_vac)) +
  geom_histogram(show.legend=FALSE) +
  facet_wrap(vars(infancy_vac), nrow=2) +
  xlab("Age in years")
```

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



```r
x <- t.test(time_length( wp$age, "years" ),
       time_length( ap$age, "years" ))

x$p.value
```

```
[1] 2.372101e-23
```

The p-value is less than 0.05, so therefore these groups are significantly different.

Obtain more data from CMI-PB

```r
specimine <- read_json("http://cmi-pb.org/api/v5_1/specimen", simplifyVector = TRUE)
ab_data <- read_json("http://cmi-pb.org/api/v5_1/plasma_ab_titer", simplifyVector= TRUE)
```

```r
head(specimine)
```

```
  specimen_id subject_id actual_day_relative_to_boost
1           1          1                           -3
2           2          1                            1
3           3          1                            3
4           4          1                            7
```

```
5           5          1                          11
6           6          1                          32
  planned_day_relative_to_boost specimen_type visit
1                             0         Blood     1
2                             1         Blood     2
3                             3         Blood     3
4                             7         Blood     4
5                            14         Blood     5
6                            30         Blood     6
```

```
head(ab_data)
```

```
  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
1 UG/ML                 2.096133
2 IU/ML                29.170000
3 IU/ML                 0.530000
4 IU/ML                 6.205949
5 IU/ML                 4.679535
6 IU/ML                 2.816431
```

I now have three tables of data from CMI-PB: `subject`, `specimine`, and `ab_data`.

I need to join these tables so I will have all the info I need to work with.

For this we will use the `inner_join()` function from the **dplyr** package.

> 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:

```
library(dplyr)

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

```
Joining with `by = join_by(subject_id)`
```

```r
head(meta)
```

```
  subject_id infancy_vac biological_sex              ethnicity  race
1          1          wP         Female Not Hispanic or Latino White
2          1          wP         Female Not Hispanic or Latino White
3          1          wP         Female Not Hispanic or Latino White
4          1          wP         Female Not Hispanic or Latino White
5          1          wP         Female Not Hispanic or Latino White
6          1          wP         Female Not Hispanic or Latino White
  year_of_birth date_of_boost      dataset        age specimen_id
1    1986-01-01    2016-09-12 2020_dataset 14312 days           1
2    1986-01-01    2016-09-12 2020_dataset 14312 days           2
3    1986-01-01    2016-09-12 2020_dataset 14312 days           3
4    1986-01-01    2016-09-12 2020_dataset 14312 days           4
5    1986-01-01    2016-09-12 2020_dataset 14312 days           5
6    1986-01-01    2016-09-12 2020_dataset 14312 days           6
  actual_day_relative_to_boost planned_day_relative_to_boost specimen_type
1                           -3                             0         Blood
2                            1                             1         Blood
3                            3                             3         Blood
4                            7                             7         Blood
5                           11                            14         Blood
6                           32                            30         Blood
  visit
1     1
2     2
3     3
4     4
5     5
6     6
```

```r
dim(meta)
```

```
[1] 1503   14
```

Now we join our `ab_data` table to `meta` so we have all the info we need about antibody leaves.

> Q10. Now using the same procedure join meta with ab_data so we can further analyze this data in terms of time of visit aP/wP, male/female etc.

```r
abdata <- inner_join(meta, ab_data)
```

```
Joining with `by = join_by(specimen_id)`
```

```r
head(abdata)
```

```
  subject_id infancy_vac biological_sex              ethnicity  race
1          1          wP         Female Not Hispanic or Latino White
2          1          wP         Female Not Hispanic or Latino White
3          1          wP         Female Not Hispanic or Latino White
4          1          wP         Female Not Hispanic or Latino White
5          1          wP         Female Not Hispanic or Latino White
6          1          wP         Female Not Hispanic or Latino White
  year_of_birth date_of_boost      dataset        age specimen_id
1    1986-01-01    2016-09-12 2020_dataset 14312 days           1
2    1986-01-01    2016-09-12 2020_dataset 14312 days           1
3    1986-01-01    2016-09-12 2020_dataset 14312 days           1
4    1986-01-01    2016-09-12 2020_dataset 14312 days           1
5    1986-01-01    2016-09-12 2020_dataset 14312 days           1
6    1986-01-01    2016-09-12 2020_dataset 14312 days           1
  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 isotype is_antigen_specific antigen        MFI MFI_normalised  unit
1     1     IgE               FALSE   Total 1110.21154       2.493425 UG/ML
2     1     IgE               FALSE   Total 2708.91616       2.493425 IU/ML
3     1     IgG                TRUE      PT   68.56614       3.736992 IU/ML
4     1     IgG                TRUE     PRN  332.12718       2.602350 IU/ML
5     1     IgG                TRUE     FHA 1887.12263      34.050956 IU/ML
6     1     IgE                TRUE     ACT    0.10000       1.000000 IU/ML
  lower_limit_of_detection
1                 2.096133
2                29.170000
3                 0.530000
4                 6.205949
5                 4.679535
6                 2.816431
```

```r
dim(abdata)
```

```
[1] 61956    21
```

Q11. How many specimens (i.e. entries in abdata) do we have for each isotype?

```r
table(abdata$isotype)
```

```

  IgE   IgG  IgG1  IgG2  IgG3  IgG4 
 6698  7265 11993 12000 12000 12000 
```

```r
table(abdata$antigen)
```

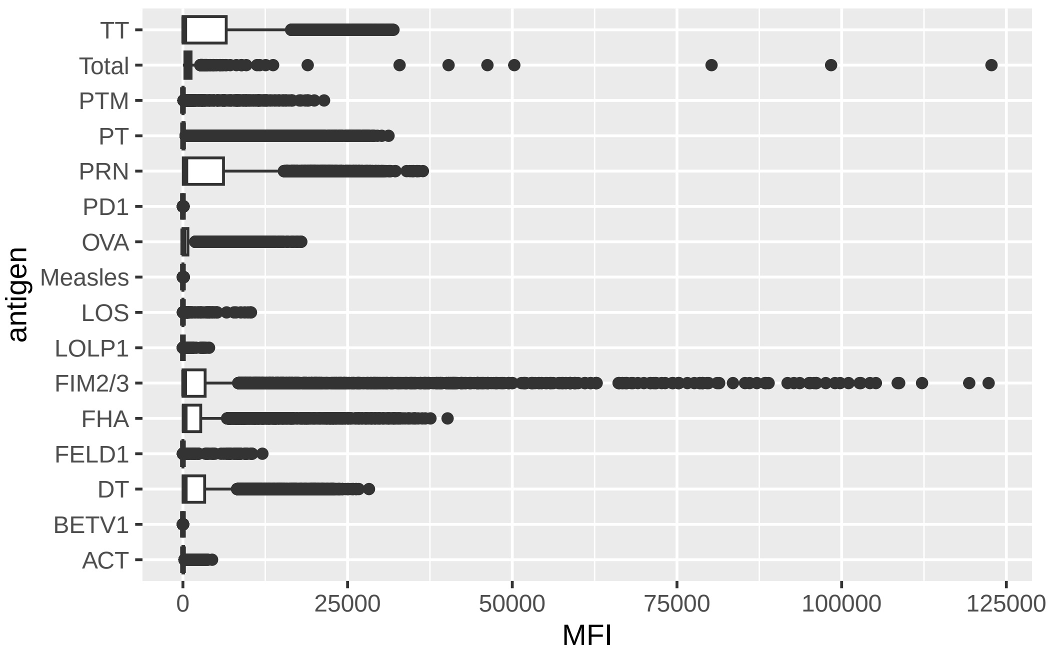
```

    ACT   BETV1      DT   FELD1     FHA  FIM2/3   LOLP1     LOS Measles     OVA 
   1970    1970    6318    1970    6712    6318    1970    1970    1970    6318 
    PD1     PRN      PT     PTM   Total      TT 
   1970    6712    6712    1970     788    6318 
```

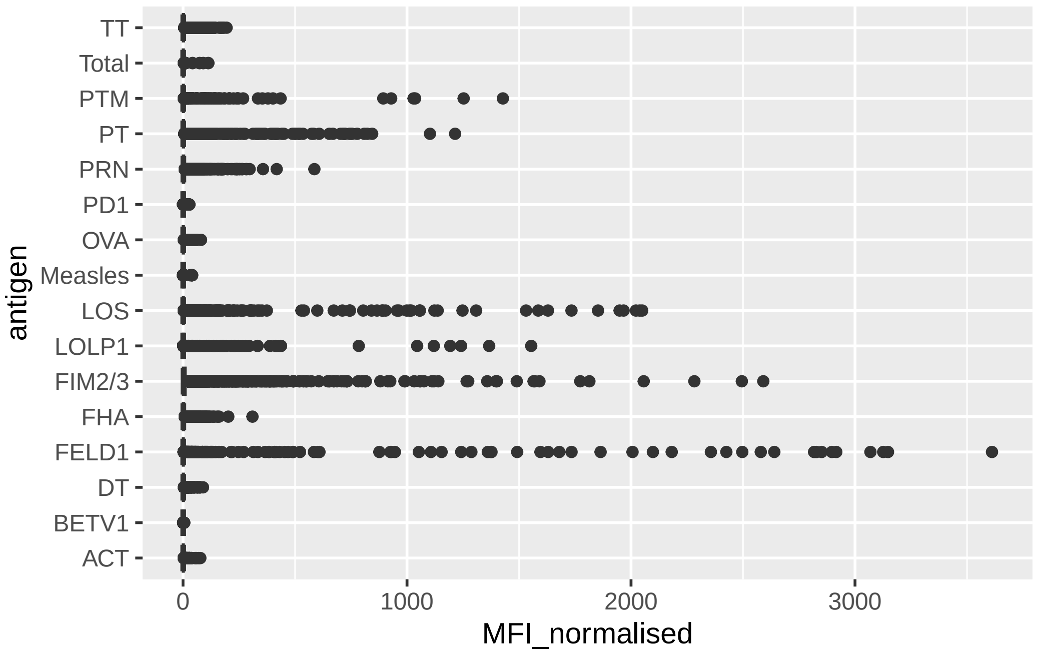
I want a plot of antigen levels across the whole dataset.

```r
ggplot(abdata) + aes(MFI, antigen) + geom_boxplot()
```

```
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_boxplot()`).
```



```r
ggplot(abdata) + aes(MFI_normalised, antigen) + geom_boxplot()
```



Antigens like FIM2/3, PT, FELD1 have quite a large range of values. Others like measles don't show much activity.

Q12. What are the different $dataset values in abdata and what do you notice about the number of rows for the most "recent" dataset?

```
table(abdata$dataset)
```

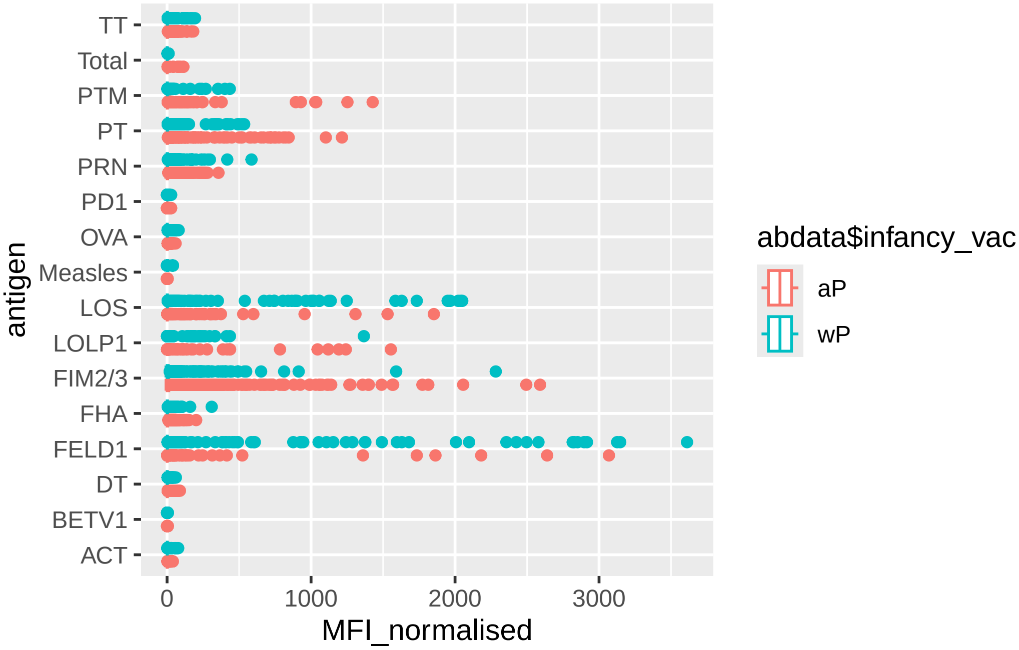
```
2020_dataset 2021_dataset 2022_dataset 2023_dataset 
       31520         8085         7301        15050 
```

In the most recent dataset in 2023, the number is almost double from the previous two years, but half of the 2020 dataset.

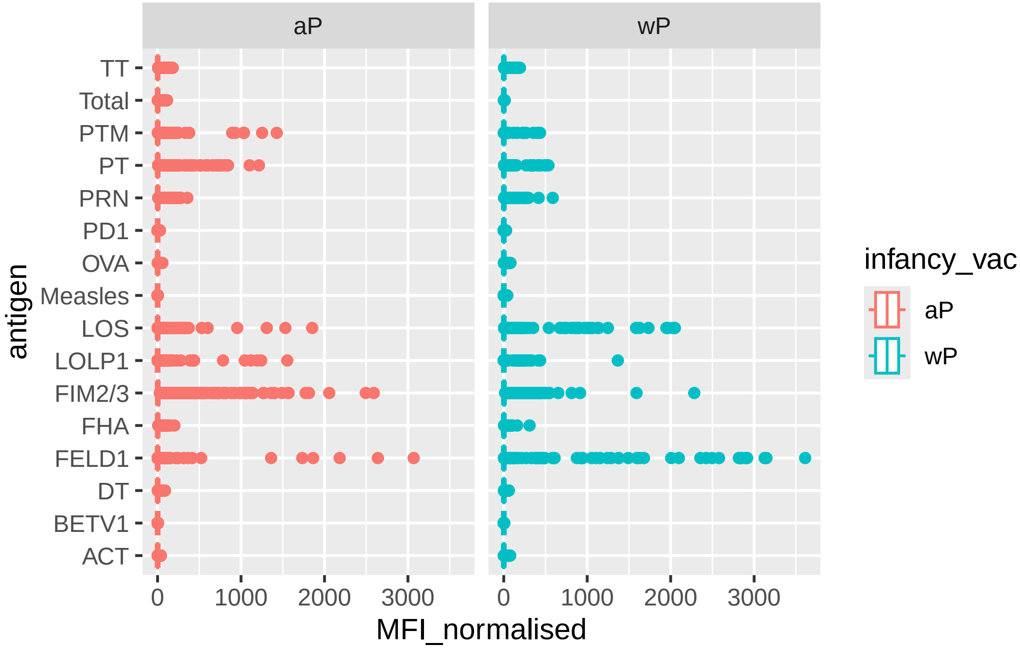
Q. Are there differences at this whole-dataset levels between aP and wP?

```
ggplot(abdata) + aes(MFI_normalised, antigen, col=abdata$infancy_vac) + geom_boxplot()
```

```
Warning: Use of `abdata$infancy_vac` is discouraged.
i Use `infancy_vac` instead.
```



```
ggplot(abdata) + aes(MFI_normalised, antigen, col=infancy_vac) + geom_boxplot() + facet_wrap
```



Examine IgG Antibody Titer Levels

For this I need to select out just isotype IgG.

```
igg <- abdata %>% filter(isotype == "IgG")
head(igg)
```

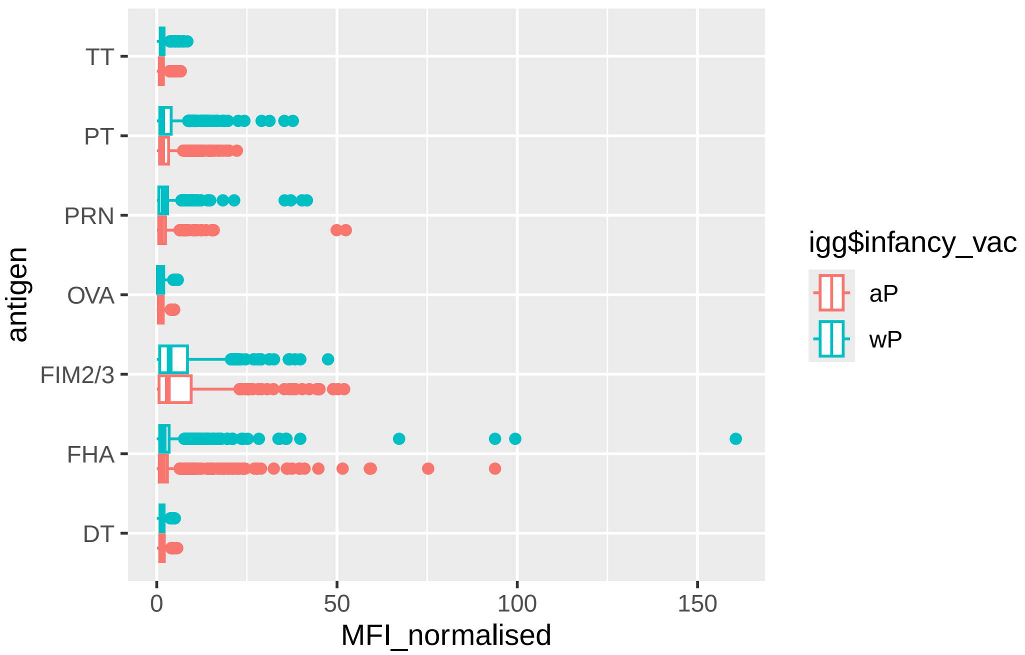
```
  subject_id infancy_vac biological_sex              ethnicity  race
1          1          wP         Female Not Hispanic or Latino White
2          1          wP         Female Not Hispanic or Latino White
3          1          wP         Female Not Hispanic or Latino White
4          1          wP         Female Not Hispanic or Latino White
5          1          wP         Female Not Hispanic or Latino White
6          1          wP         Female Not Hispanic or Latino White
  year_of_birth date_of_boost      dataset        age specimen_id
1    1986-01-01    2016-09-12 2020_dataset 14312 days           1
2    1986-01-01    2016-09-12 2020_dataset 14312 days           1
3    1986-01-01    2016-09-12 2020_dataset 14312 days           1
4    1986-01-01    2016-09-12 2020_dataset 14312 days           2
```

```
5    1986-01-01    2016-09-12 2020_dataset 14312 days           2
6    1986-01-01    2016-09-12 2020_dataset 14312 days           2
  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                            1                             1         Blood
5                            1                             1         Blood
6                            1                             1         Blood
  visit isotype is_antigen_specific antigen        MFI MFI_normalised  unit
1     1     IgG                TRUE      PT   68.56614       3.736992 IU/ML
2     1     IgG                TRUE     PRN  332.12718       2.602350 IU/ML
3     1     IgG                TRUE     FHA 1887.12263      34.050956 IU/ML
4     2     IgG                TRUE      PT   41.38442       2.255534 IU/ML
5     2     IgG                TRUE     PRN  174.89761       1.370393 IU/ML
6     2     IgG                TRUE     FHA  246.00957       4.438960 IU/ML
  lower_limit_of_detection
1                 0.530000
2                 6.205949
3                 4.679535
4                 0.530000
5                 6.205949
6                 4.679535
```

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

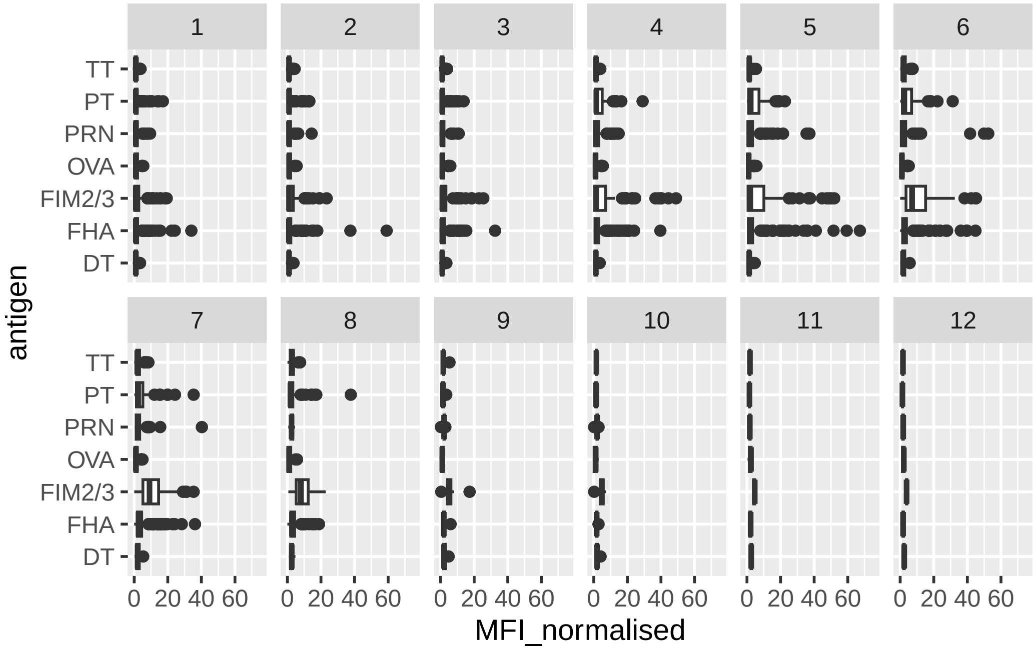
```
ggplot(igg) + aes(MFI_normalised, antigen, col=igg$infancy_vac) + geom_boxplot()
```

```
Warning: Use of `igg$infancy_vac` is discouraged.
i Use `infancy_vac` instead.
```



```r
ggplot(igg) +
  aes(MFI_normalised, antigen) +
  geom_boxplot() +
    xlim(0,75) +
  facet_wrap(vars(visit), nrow=2)
```

```
Warning: Removed 5 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



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

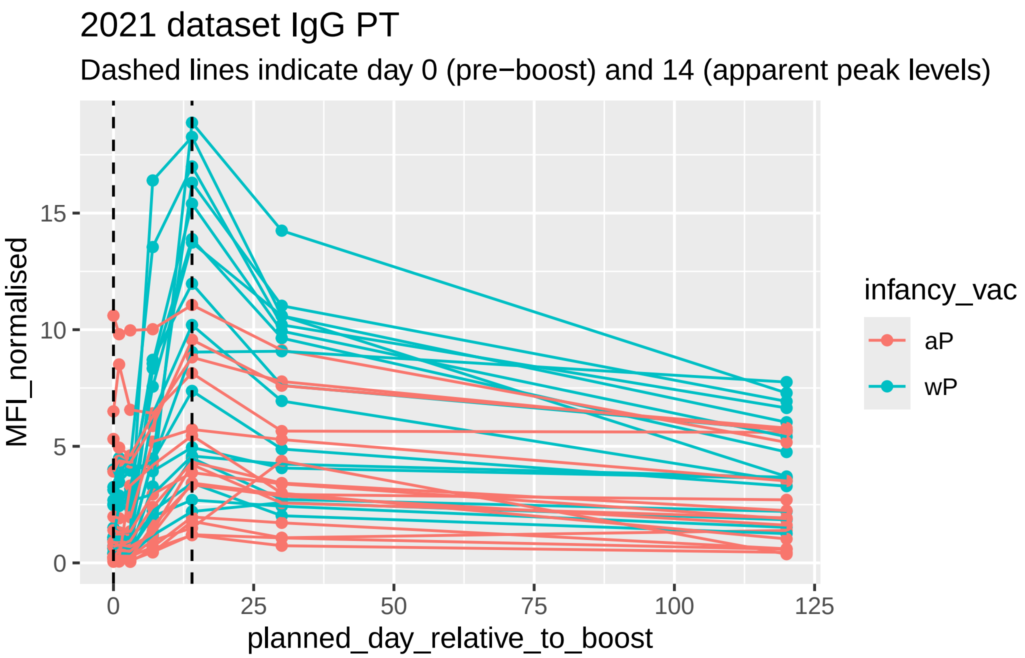
We see that FIM2/3 and PT have the highest differences across the IgG antibody titers over time. This is because PT is the pertussis toxin, which is a virulence factor produces by the bacterium. FIM2/3 relates to the Fimbriae on the pertussis bacterium. These two are part of the whole-cell vaccine components and therefore will be used to target bacteruim in the human body. Since these are present on the bacteruim, the antibosies will be reconzing them more over time since they will be present during infection.

Digging in further to look at the time course of IgG isotype PT antigen leaves across aP and wP individuals:

```r
#Filter to include 2021 data only
abdata.21 <- abdata |> filter(dataset == "2021_dataset")

#Filter to look at IgG PT data only
pt.igg <- abdata.21 |>
  filter(isotype == "IgG",  antigen == "PT")

#Plot and color by infancy_vac(wP and aP)
  ggplot(pt.igg) +
    aes(x=planned_day_relative_to_boost,
        y=MFI_normalised,
        col=infancy_vac,
        group=subject_id) +
    geom_point() +
    geom_line() +
    geom_vline(xintercept=0, linetype="dashed") +
    geom_vline(xintercept=14, linetype="dashed") +
  labs(title="2021 dataset IgG PT",
       subtitle = "Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)")
```


2021 dataset IgG PT
Dashed lines indicate day 0 (pre−boost) and 14 (apparent peak levels)
MFI_normalised
planned_day_relative_to_boost
infancy_vac
aP
wP