Lab 18: Pertussis Mini Project

Jessica Le (PID: A17321021)

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

The CDC tracks Pertussis cases around the US. http://tinyurl.com/pertussiscdc

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

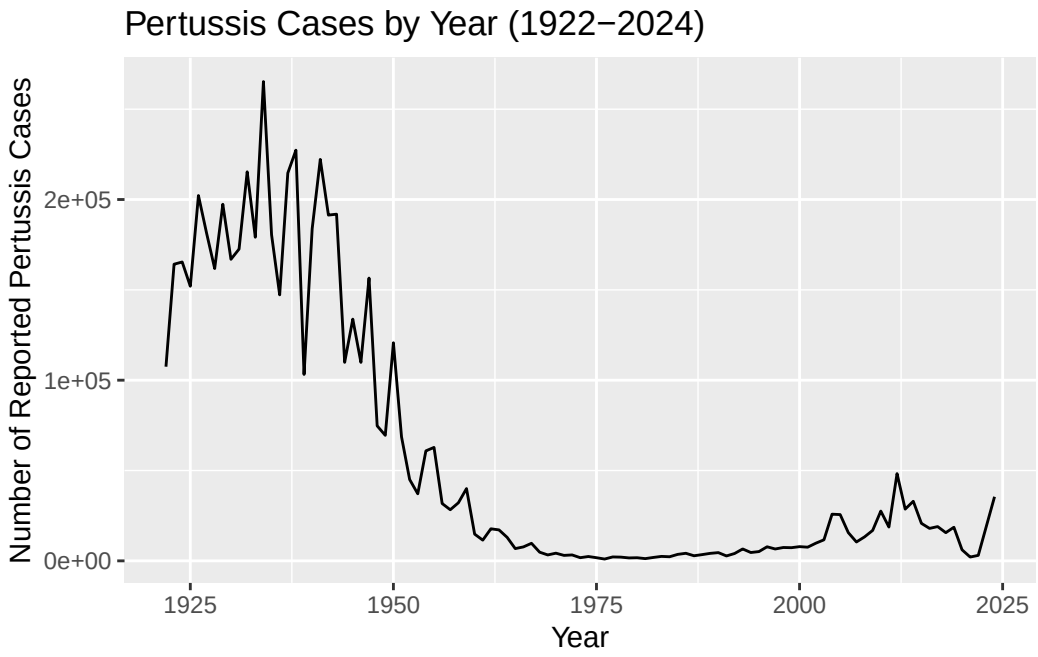
```
head(cdc)
```

```
  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.

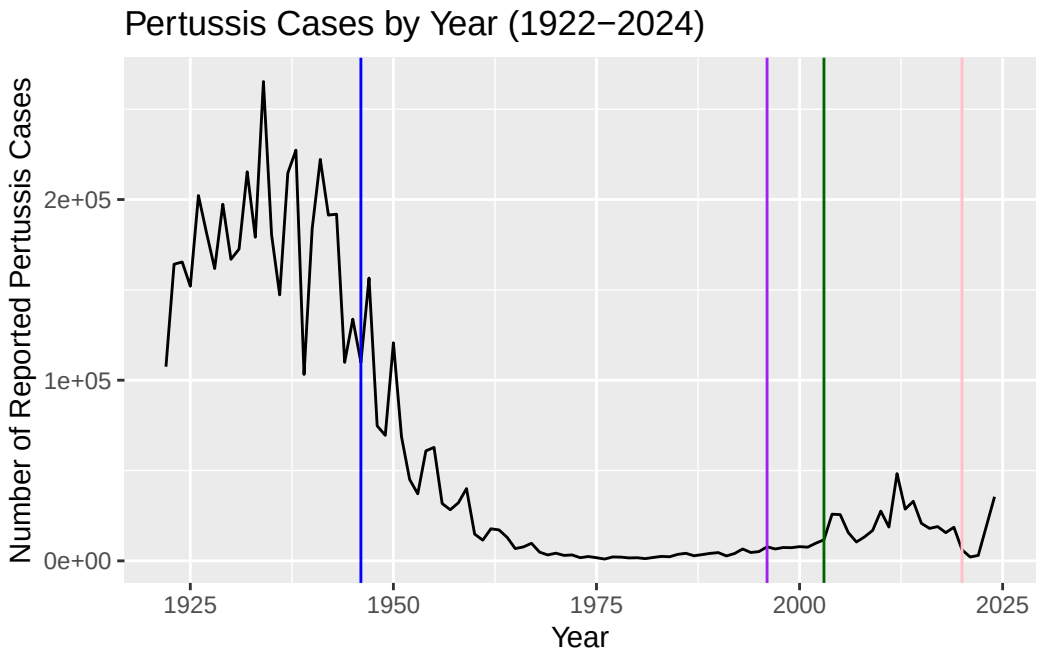
```
library(ggplot2)

ggplot(cdc) +
  aes(year, cases) +
  geom_line() +
  labs(title="Pertussis Cases by Year (1922-2024)", x="Year", y="Number of Reported Pertussis
```



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_line() +
  geom_vline(xintercept = 1946, col="blue") +
  geom_vline(xintercept = 1996, col="purple") +
  geom_vline(xintercept=2020, col="pink") +
  geom_vline(xintercept=2003, col="darkgreen") +
  labs(title="Pertussis Cases by Year (1922-2024)", x="Year", y="Number of Reported Pertussis
```



There were high cases numbers before the first wP (whole-cell) vaccine roll out in 1946 where a rapid decline in cases numbers is observed until 2004 when we have our first large-scale outbreaks of pertussis again. There is also notable COVID decrease in the number of reported outbreaks, and a recent rapid rise in cases has been reported in the recent years.

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

There appears to be an increase in the number of reported pertussis cases after the introduction of the aP vaccine. A possible explanation includes the pertussis bacteria evolving to become resistant to the vaccine; therefore, vaccination is not helping to reduce the number of reported cases since people can still be infected due to the resistant bacteria.

Computational Models of Immunity Pertussis Boost (CMI-PB)

> Q. What is different about the immune response to infection if you had an older wP vaccine vs the newer aP vaccine?

No definitive answer can be provided right now regarding the difference in immune response to infection between the different vaccines.

The CMI-PB project aims to address this key question: what is different 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.

```
library(jsonlite)
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 individuals "subjects" are in this dataset?

```
nrow(subject)
```

```
[1] 172
```

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

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

```
aP wP
87 85
```

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

```
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…)?

```
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
```

This is not representative of the US population but it is the biggest dataset of its type so let's see what we can learn…

Side-Note: Working with dates

```
library(lubridate)
```

```
Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

Today's date is…

```
today()
```

```
[1] "2025-03-08"
```

The number of days that have passed since new year 2000...

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

```
Time difference of 9198 days
```

The time difference of 9198 days in years is...

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

```
[1] 25.18275
```

> Q7. Determine (i) the average age of wP individuals, (ii) the average age of aP individuals; and (iii) are they significantly different?

Use dplyr's filter() function to limit ourselves to a particular subset of subjects to examine the 6 number summary of their age in years:

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

```
Time differences in days
  [1] 14311 20886 15407 13581 12485 13581 16137 14676 10659 15772 14311 15772
 [13] 10293 11754 13215 13946 16503 10293 11389 16137 15407 14676 12485 12120
 [25] 13581 15407 10293 15772 10293 13581 13215 10293 12850 15407 12485 10293
 [37]  9928 10293 14676 11389 14676 10293  9928  9928 10293  9928 10659  9928
 [49] 10293 10293 10293  9928  9928 10293 10293 10293 10659 10293 10293 10293
 [61] 13946 11754 11024 11754 12850 17964 19425 19425 12850  9928  9928 12485
 [73] 11024 11024  9928  9928 13581 11754 13946 12120 11754  9928  9563 10293
 [85]  9198  9928  9198  9198 10293  9563  9928  9198 10659  9563  9928  9198
 [97] 14311 11754  9563  8832  8102  8102 11389 13215 11389 10659  9928 11024
[109] 13215 10293 10659 10659 10659 12850  8467  9198 11389  9928  9928 11024
[121]  9198  9563 10659  9198 11754 11754 10659 11389 12485 10659  9928 11024
[133] 10293 12850 11024 11024  9928  9198 11754  8832 10659 12485  8102  9563
[145]  8467 12120  9198 13581 12485 12485 12120 11024  9928 10293 10293  8832
[157] 10293  9198 11389 10659 11754  9563 11754 12485 11754  8832 10293 12485
[169]  8102 12120  8102 14311
```

```
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
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
```

```
# wP
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
```

The average age of aP individuals is 27 and 36 for wP individuals. The difference in their average age is 9 and does appear to be significantly different.

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

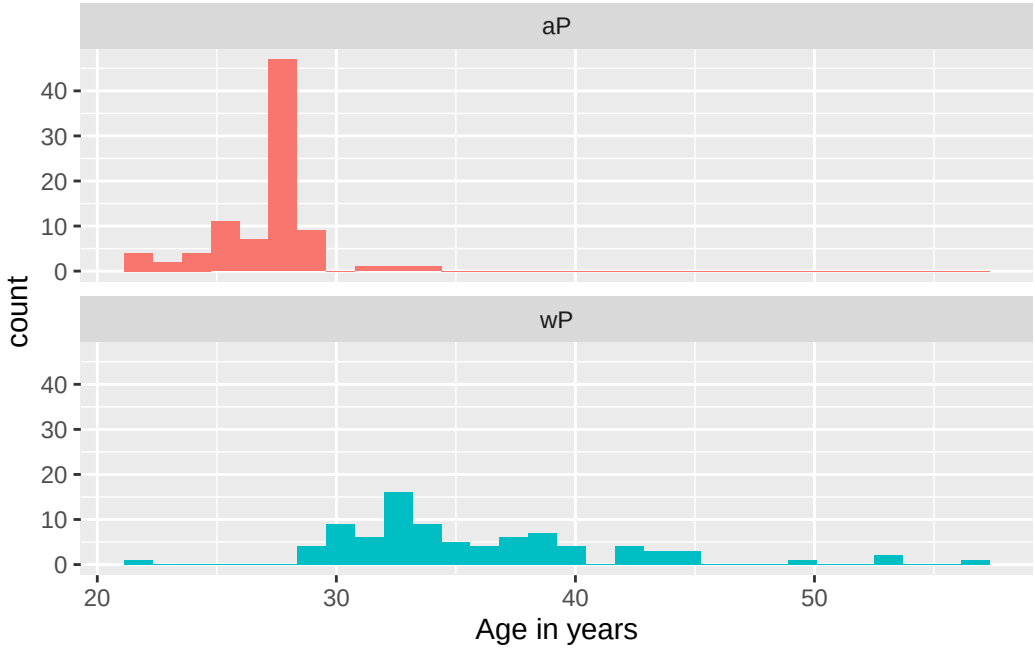
```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

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

Q9. With the help of a faceted boxplot or histogram, 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) +
  xlab("Age in years")
```

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



Yes, average age of the wP and aP individuals does appear different in the histograms. We can confirm whether or not they are statisticaly significant by calculating a p-value.

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

x$p.value
```

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

Because the calculated p-value is less than 0.05, it can be concluded that the average age between wP and aP individuals are statistically significant.

Joining multiple tables

Obtain more data from CMI-PB to explore more about the data...

```
specimen <- read_json("https://www.cmi-pb.org/api/v5_1/specimen",
                      simplifyVector=T)
ab_data <- read_json("https://www.cmi-pb.org/api/v5_1/plasma_ab_titer",
                      simplifyVector=T)
```

```
head(specimen)
```

```
  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 3 tables of data from CMI-PB: `subject`, `specimen`, 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, specimen)
```

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

```
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 14311 days           1
2    1986-01-01    2016-09-12 2020_dataset 14311 days           2
3    1986-01-01    2016-09-12 2020_dataset 14311 days           3
4    1986-01-01    2016-09-12 2020_dataset 14311 days           4
5    1986-01-01    2016-09-12 2020_dataset 14311 days           5
6    1986-01-01    2016-09-12 2020_dataset 14311 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
```

```
dim(subject)
```

```
[1] 172   9
```

```
dim(specimen)
```

```
[1] 1503    6
```

```
dim(meta)
```

```
[1] 1503   14
```

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

> 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(meta, ab_data)
```

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

```
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 14311 days           1
2    1986-01-01    2016-09-12 2020_dataset 14311 days           1
3    1986-01-01    2016-09-12 2020_dataset 14311 days           1
4    1986-01-01    2016-09-12 2020_dataset 14311 days           1
5    1986-01-01    2016-09-12 2020_dataset 14311 days           1
6    1986-01-01    2016-09-12 2020_dataset 14311 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
```

Q. How many different antibody isotypes are there in this dataset?

```
length(abdata$isotype)
```

```
[1] 61956
```

There are 61956 total different antibody istoypes in the dataset.

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

To find the amount of each antibody isotype:

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

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

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
```

The number of rows for the most recent dataset of 2023 has significantly increased compared to the previous decrease in the years of 2021 and 2022 from 2020.

Q. How many different antigens are in this dataset?

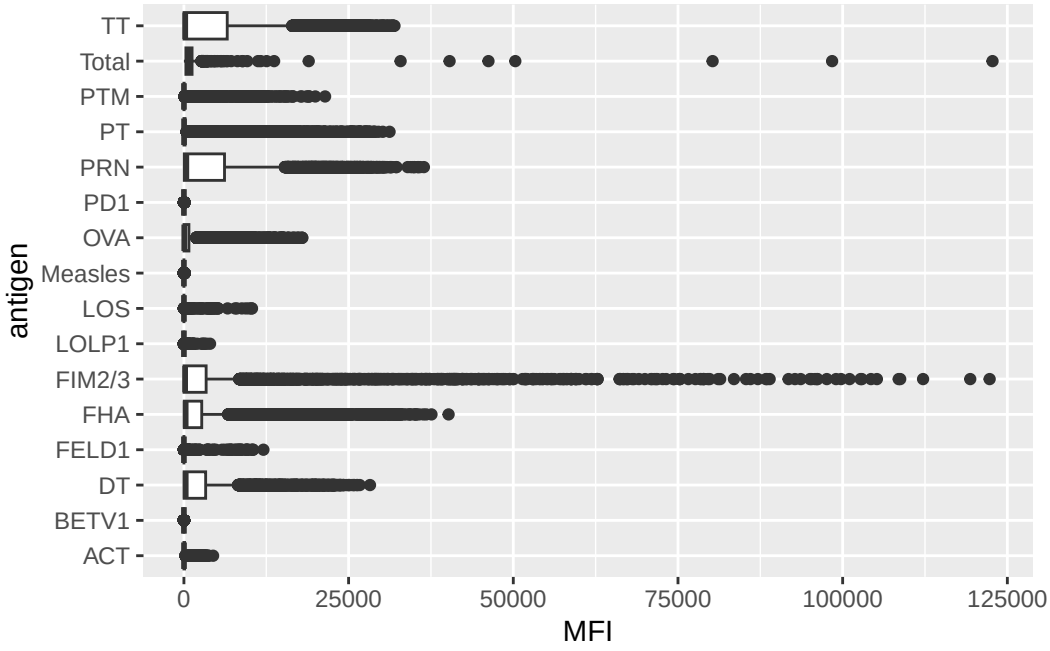
```
table(abdata$antigen)
```

```
    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
```

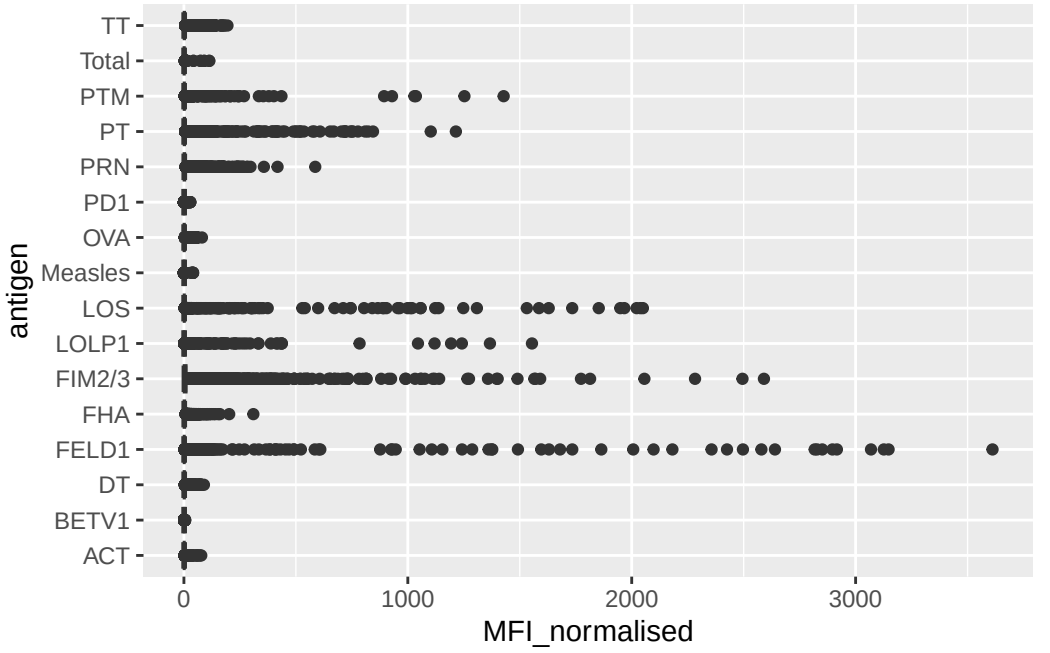
I want a plot of antigen levels across the whole dataset.

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

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



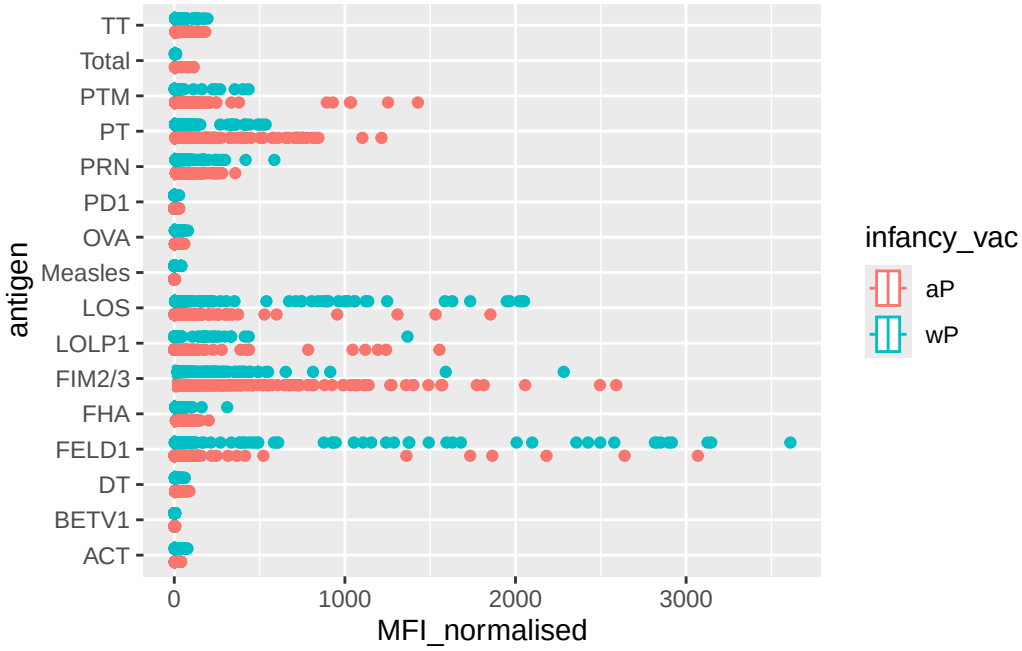

```
ggplot(abdata) +
  aes(MFI_normalised, antigen) +
  geom_boxplot()
```



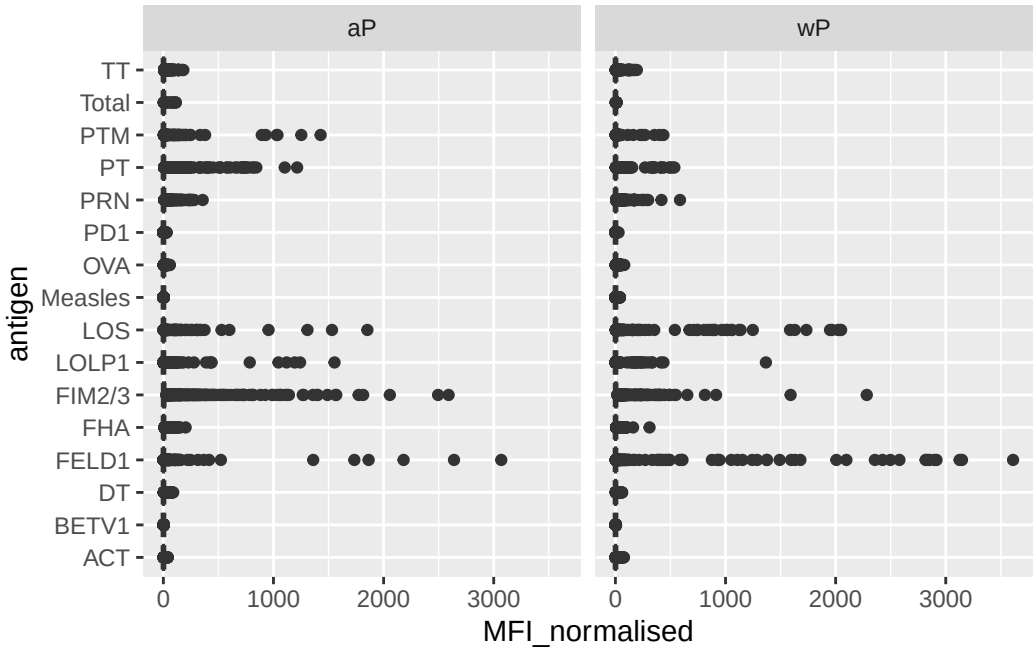
Antigens like FIM2/3, PT, and FELD1 have quite a large range of values. Others like Measles don't show much activity. Antigens with high range of values are those present in either the wP or aP vaccine.

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

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




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



Examine IgG Ab titer levels

For this I need to select out just isotype IgG.

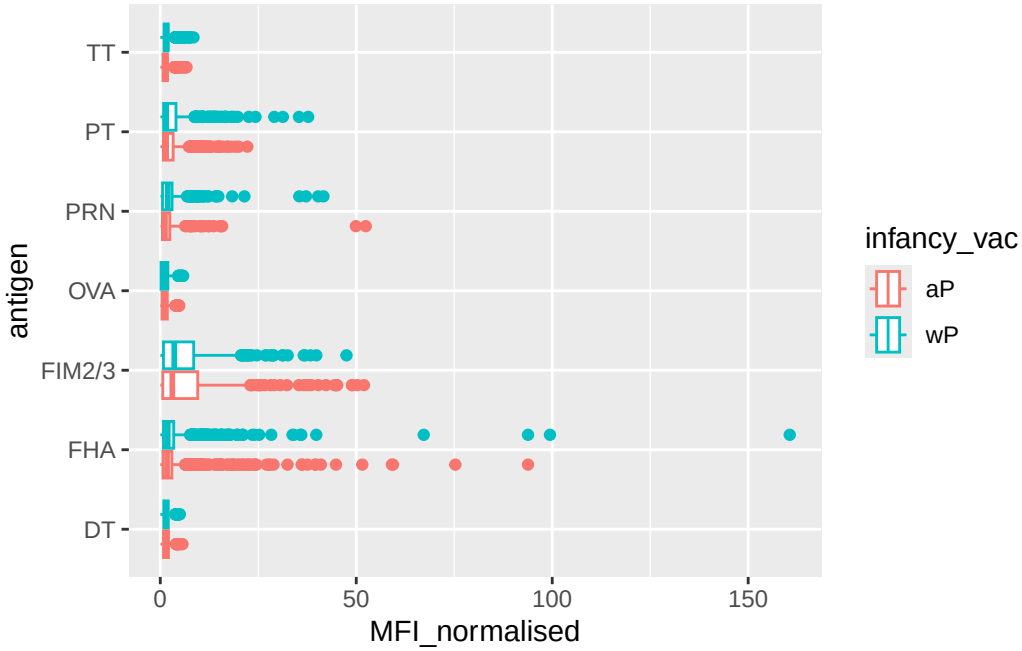
```
igg <- abdata %>% filter(isotype == "IgG")
head(igg)
```

```
  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 14311 days           1
2    1986-01-01    2016-09-12 2020_dataset 14311 days           1
3    1986-01-01    2016-09-12 2020_dataset 14311 days           1
4    1986-01-01    2016-09-12 2020_dataset 14311 days           2
5    1986-01-01    2016-09-12 2020_dataset 14311 days           2
6    1986-01-01    2016-09-12 2020_dataset 14311 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
```

An overview boxplot:

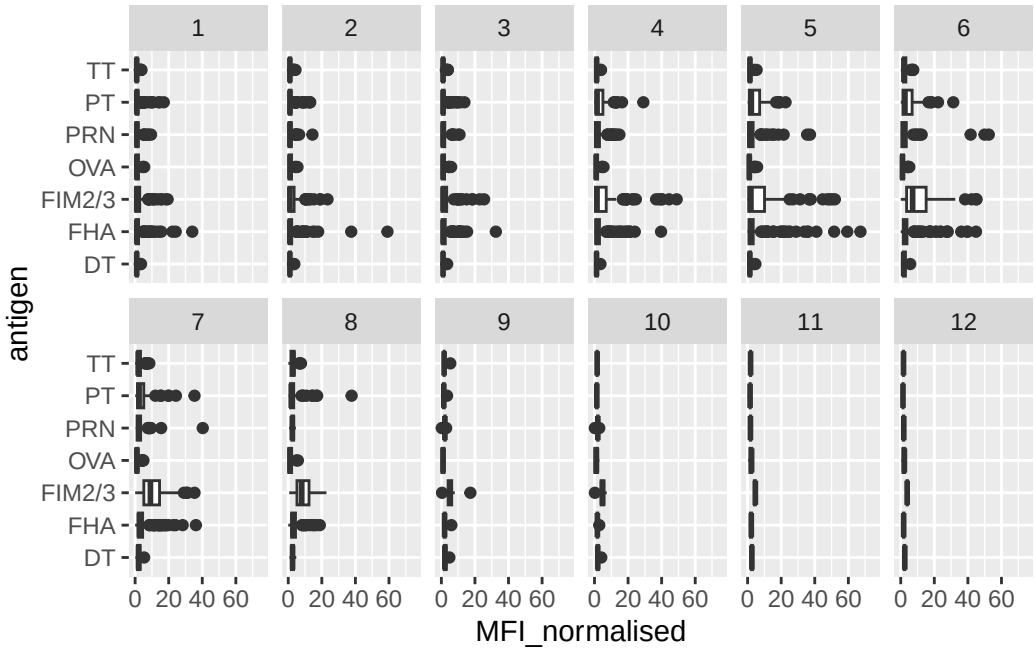
```
ggplot(igg) +
  aes(MFI_normalised, antigen, col=infancy_vac) +
  geom_boxplot()
```



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

```
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?

FIM2/3, PT, and FHA are antigens that show differences in the level of IgG antibody titers recognizing them over time. They are components of the pertussis vaccine; therefore, will have different levels of IgG antibody titers recognition over time as they will attempt to provide protection against the different strains of pertussis that evolve over time.

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

```
## 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 vs 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)")
```