class18: Investigating Pertussis Resurgence

Shivani

Pertussis (more commonly known as whooping cough) is a highly contagious respiratory disease caused by the bacterium Bordetella pertussis.

The United States Centers for Disease Control and Prevention (CDC) has been compiling reported pertussis case numbers since 1922 in their National Notifiable Diseases Surveillance System (NNDSS). We can view this data on the CDC website here: https://www.cdc.gov/pertussis/surv-reporting/cases-by-year.html

datapasta package is used to copy the data from the website and paste it in R.

```
head(cdc)
```

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

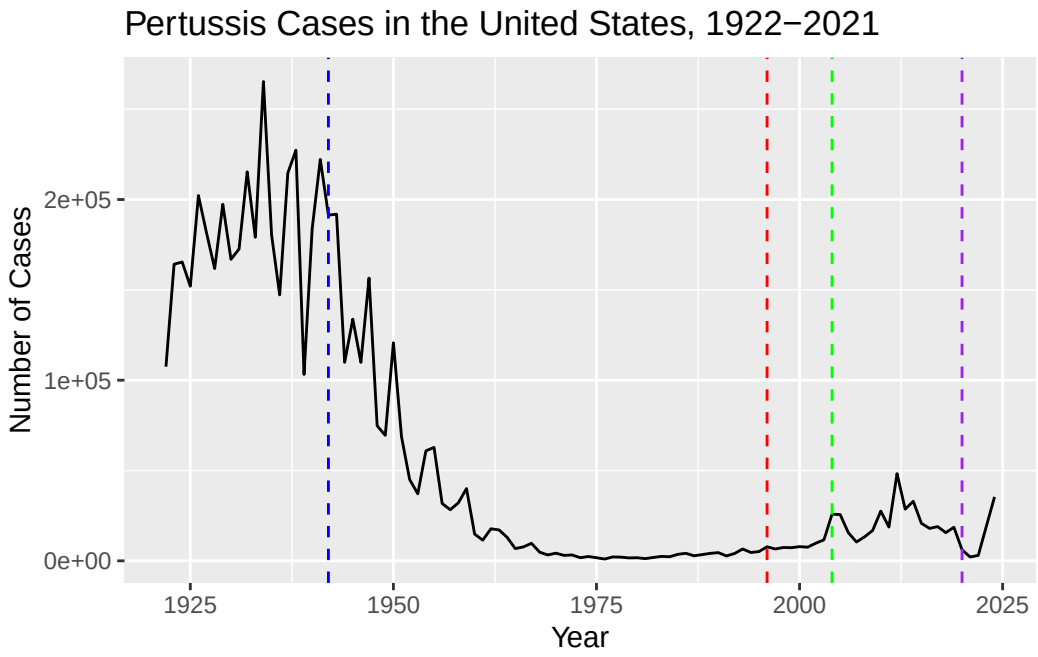
> Q.1 With the help of the R "addin" package datapasta assign the CDC pertussis case number data to a data frame called cdc and use ggplot to make a plot of cases numbers over time.

```
library(ggplot2)

ggplot(cdc, aes(x = year, y = cases)) +
  geom_line() +
  labs(title = "Pertussis Cases in the United States, 1922-2021",
       x = "Year",
       y = "Number of Cases")
```

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(x = year, y = cases)) +
  geom_line() +
  geom_vline(xintercept = 1942, linetype = "dashed", col="blue") +
  geom_vline(xintercept = 1996, linetype = "dashed", col="red") +
  geom_vline(xintercept = 2020, linetype = "dashed", col="purple") +
  geom_vline(xintercept = 2004, linetype = "dashed", col="green") +
  labs(title = "Pertussis Cases in the United States, 1922-2021",
       x = "Year",
       y = "Number of Cases")
```



There were many cases pre 1946 (before wP vaccine), with rapid decrease in cases through 1970s and on to 2004 when our first widespread outbreak occurred again. There is **waning efficacy** of the aP vaccine after ~10 years, or faster than the wP vaccine.

Mounting evidence indicates that the acellular pertussis (aP) vaccine is less effective than the whole-cell pertussis (wP) vaccine.

Enter the CMI-PB project

Computational Models of Immunity Pertussis Boost

```
library(jsonlite)

subject <- read_json("http://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
```

Q3. How many subjects are in the dataset?

```
nrow(subject)
```

```
[1] 172
```

172 individuals!

Q4. How many aP and wP infancy vaccinated subjects are in the dataset?

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

```
aP wP
87 85
```

87 aP and 85 wP

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

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

```
Female   Male
   112     60
```

There are 112 females and 60 males.

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
```

see table above.

Q is this represenative of the US population?

no, its from UCSD students.

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

```
library(lubridate)
```

```
Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

```
subject$age <- time_length(today() - ymd(subject$year_of_birth), "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?

```
mean(subject$age[subject$infancy_vac == "wP"])
```

```
[1] 35.82607
```

```
mean(subject$age[subject$infancy_vac == "aP"])
```

```
[1] 27.07536
```

the individuals who got wP are significantly older than those who got aP.

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

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

```
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_titer)
```

```
  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
```

Q9b. 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)
```

```
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(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 39.17864           1
2    1986-01-01    2016-09-12 2020_dataset 39.17864           2
```

```
3    1986-01-01    2016-09-12 2020_dataset 39.17864           3
4    1986-01-01    2016-09-12 2020_dataset 39.17864           4
5    1986-01-01    2016-09-12 2020_dataset 39.17864           5
6    1986-01-01    2016-09-12 2020_dataset 39.17864           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
```

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.

```
ab_data <- inner_join(meta, ab_titer)
```

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

```
head(ab_data)
```

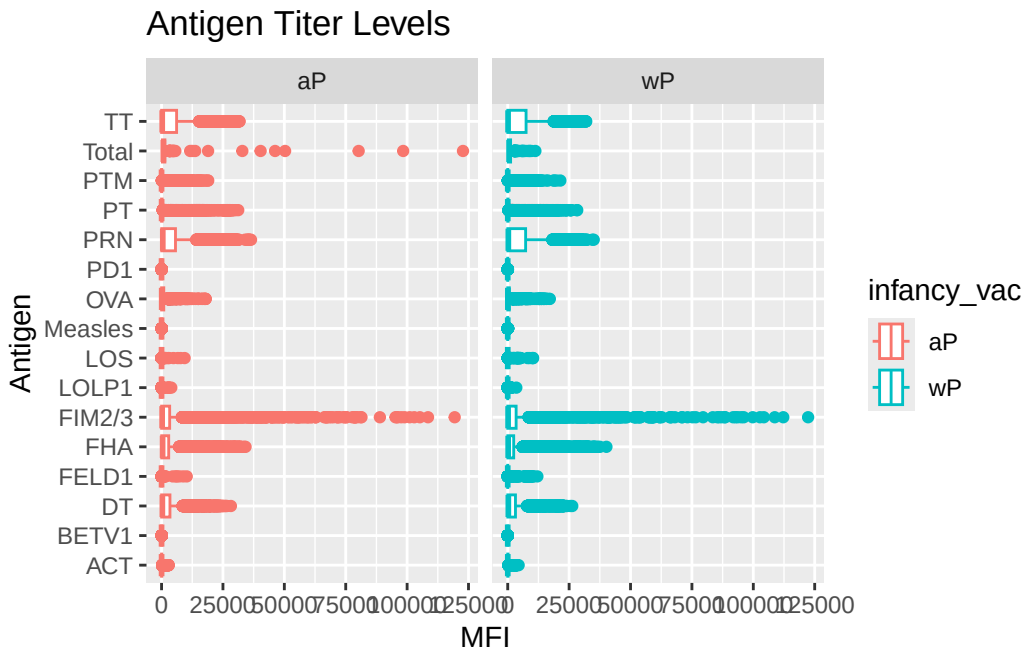
```
  subject_id infancy_vac biological_sex              ethnicity  race
1          1          wP         Female Not Hispanic or Latino White
2          1          wP         Female Not Hispanic or Latino White
3          1          wP         Female Not Hispanic or Latino White
4          1          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 39.17864           1
2    1986-01-01    2016-09-12 2020_dataset 39.17864           1
3    1986-01-01    2016-09-12 2020_dataset 39.17864           1
4    1986-01-01    2016-09-12 2020_dataset 39.17864           1
5    1986-01-01    2016-09-12 2020_dataset 39.17864           1
```

```
6    1986-01-01    2016-09-12 2020_dataset 39.17864           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
```

Q9a. With the help of a faceted boxplot (see below), do you think these two groups are significantly different?

```
ggplot(ab_data, aes(x = MFI, y = antigen, colour = infancy_vac)) +
  geom_boxplot() +
  facet_wrap(~infancy_vac) +
  labs(title = "Antigen Titer Levels",
       x = "MFI",
       y = "Antigen")
```

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

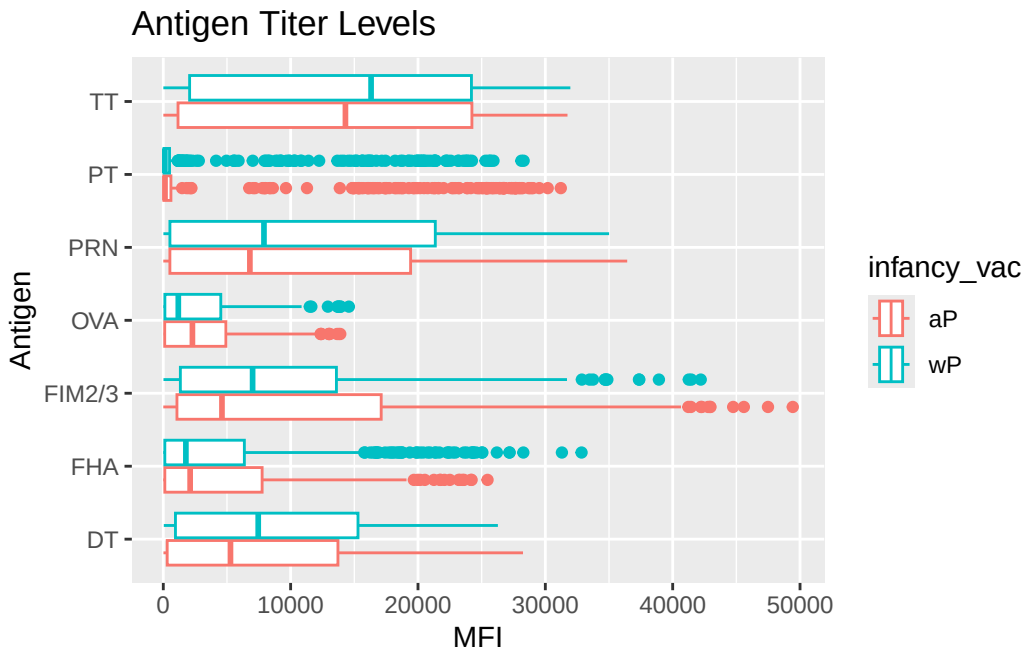


no visible difference between the two groups.

```
igg <- ab_data |>
  filter(isotype == "IgG")
```

lets make the boxplot above for just igg:

```
ggplot(igg, aes(x = MFI, y = antigen, colour = infancy_vac)) +
  geom_boxplot() +
  labs(title = "Antigen Titer Levels",
       x = "MFI",
       y = "Antigen")
```



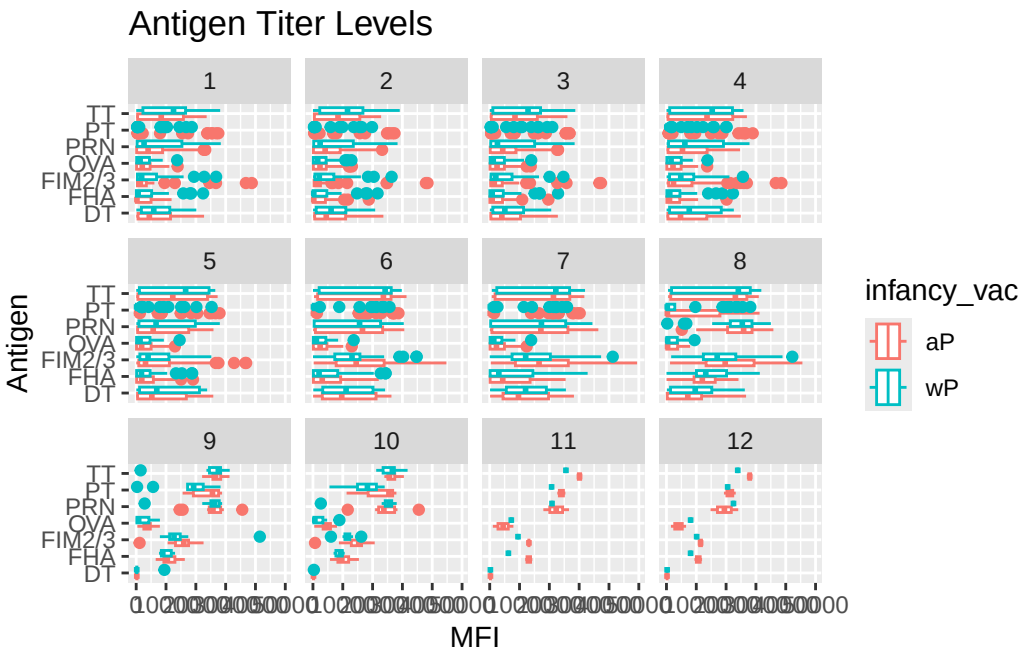

```
head(igg)
```

```
  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 39.17864           1
2    1986-01-01    2016-09-12 2020_dataset 39.17864           1
3    1986-01-01    2016-09-12 2020_dataset 39.17864           1
4    1986-01-01    2016-09-12 2020_dataset 39.17864           2
5    1986-01-01    2016-09-12 2020_dataset 39.17864           2
6    1986-01-01    2016-09-12 2020_dataset 39.17864           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
```

boxplot faceted by visit:

```
ggplot(igg, aes(x = MFI, y = antigen, colour = infancy_vac)) +
  geom_boxplot() +
  facet_wrap(~visit) +
  labs(title = "Antigen Titer Levels",
       x = "MFI",
       y = "Antigen")
```



Q11. How many specimens (i.e. entries in abdata) do we have for each isotype? How many different antibody isotypes are measured in this dataset?

```
table(ab_data$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(ab_data$dataset)
```

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

the most recent 2023 dataset has more rows than 2021 and 2022.

How many different antigens are measured in this dataset?

```
table(ab_data$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
```

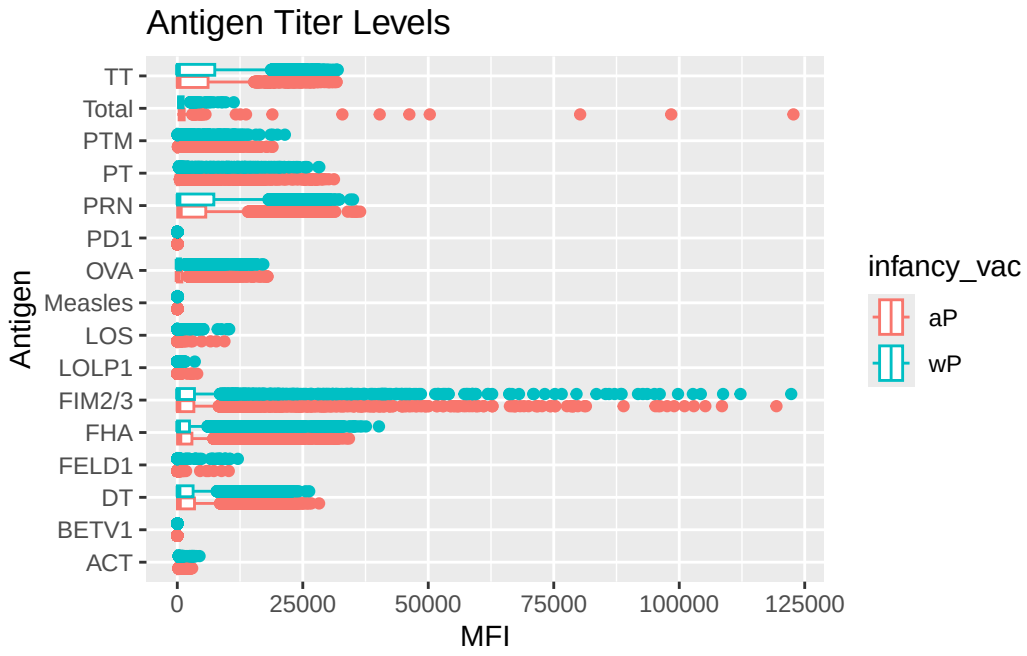
```
dim(ab_data)
```

```
[1] 61956    21
```

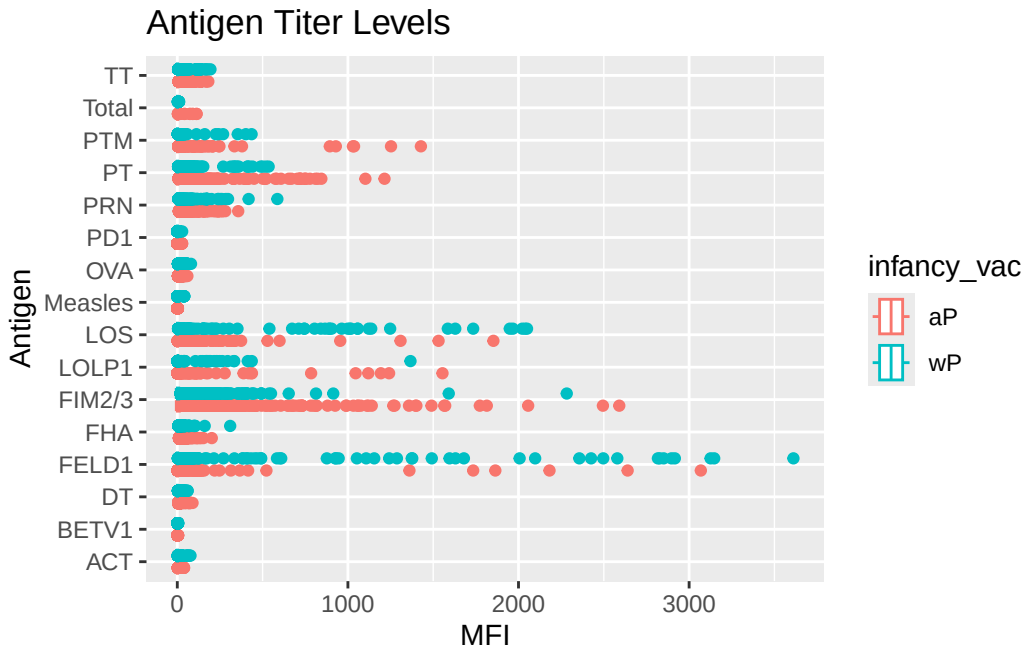
Q13. Complete the following code to make a summary boxplot of Ab titer levels for all antigens:

```
ggplot(ab_data, aes(x = MFI, y = antigen, colour = infancy_vac)) +
  geom_boxplot() +
  labs(title = "Antigen Titer Levels",
       x = "MFI",
       y = "Antigen")
```

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



```
ggplot(ab_data, aes(x = MFI_normalised, y = antigen, colour = infancy_vac)) +
  geom_boxplot() +
  labs(title = "Antigen Titer Levels",
       x = "MFI",
       y = "Antigen")
```

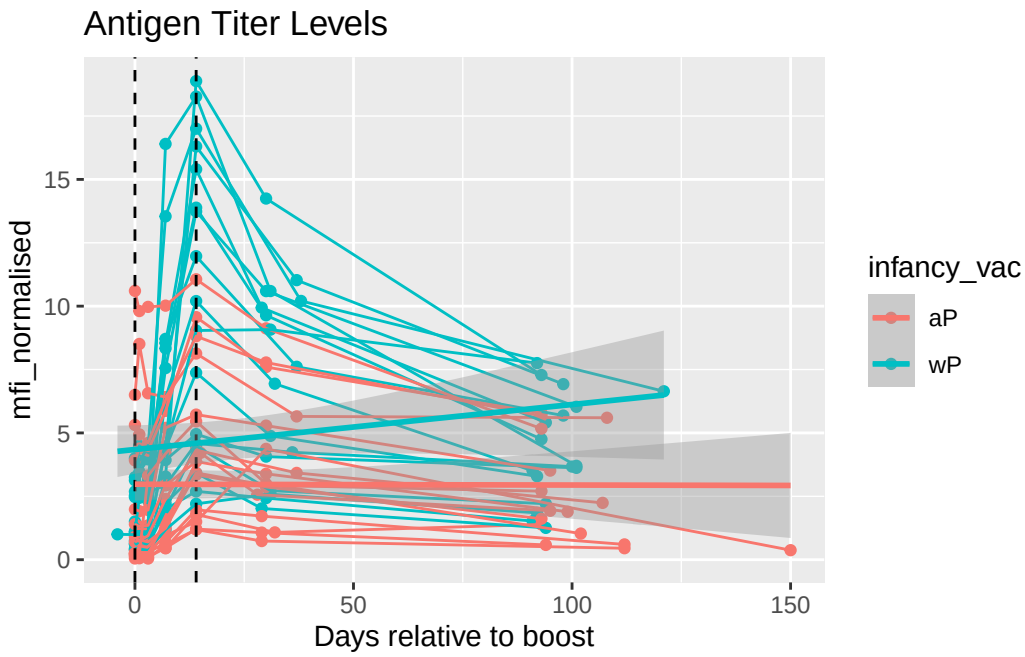


Q14. Antigen levels time-course plot for PT and igg levels over time:

```
# filter to focus on PT and IgG
pt_igg <- ab_data |>
  filter(isotype == "IgG", antigen == "PT", dataset == "2021_dataset")
```

```
ggplot(pt_igg, aes(x = actual_day_relative_to_boost, y = MFI_normalised, colour = infancy_vac
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 14, linetype = "dashed") +
  geom_vline(xintercept = 0, linetype = "dashed") +
  geom_smooth(aes(group = infancy_vac), method = "glm", se = TRUE) +
  labs(title = "Antigen Titer Levels",
       x = "Days relative to boost",
       y = "mfi_normalised")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



overall levels higher for wP than aP, but peak at 14 days post boost for both groups.