Class 19: Investigating Pertussis Resurgence

Katelyn Wei (PID: A16682595)

Investigating Pertussis Cases By Year

B. Pertussis is a severe bacterial lung infection also known as whooping cough.

We will begin by investigating the number of Pertussis cases by year in the US.

This data is available on this CDC page.

We'll use the addin DataPasta to import this data into R. It allows you to paste data you've copied into R as r objects like data frames and tables.

Let's have a quick look at this data.frame:

```
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()
```
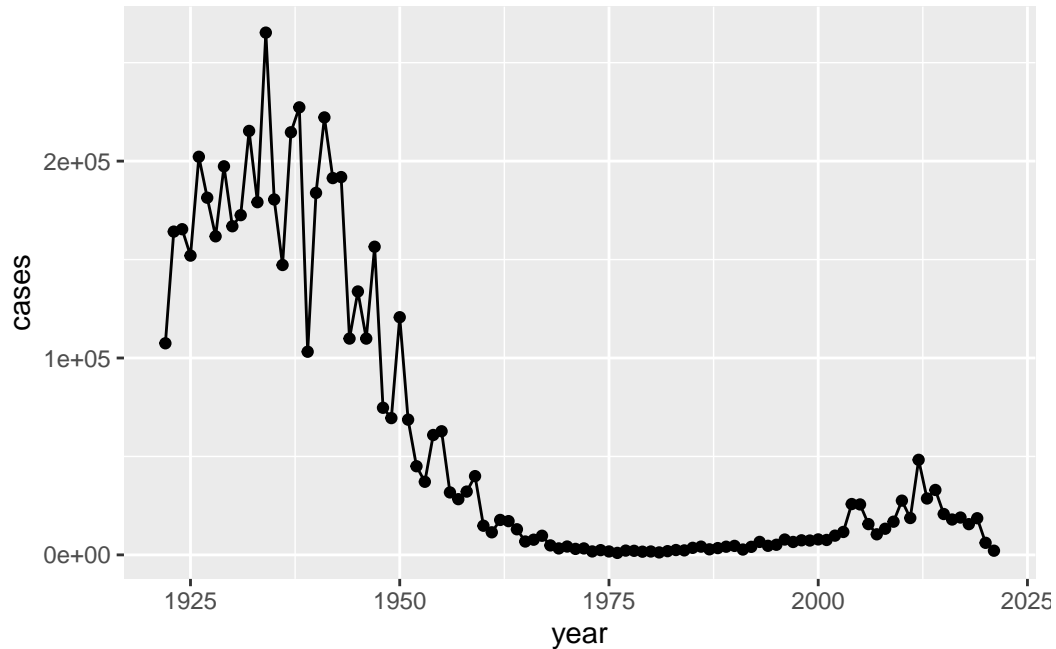
Q2. Using the ggplot geom_vline() function add lines to your previous plot for the 1946 introduction of the wP vaccine and the 1996 switch to aP vaccine (see example in the hint below). What do you notice?

```
ggplot(cdc, aes(year, cases)) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 1946, linetype = "dashed", col = "blue") +
  geom_vline(xintercept = 1996, linetype = "dashed", col = "red") +
  geom_vline(xintercept = 2020, linetype = "dashed", col = "darkgrey")
```
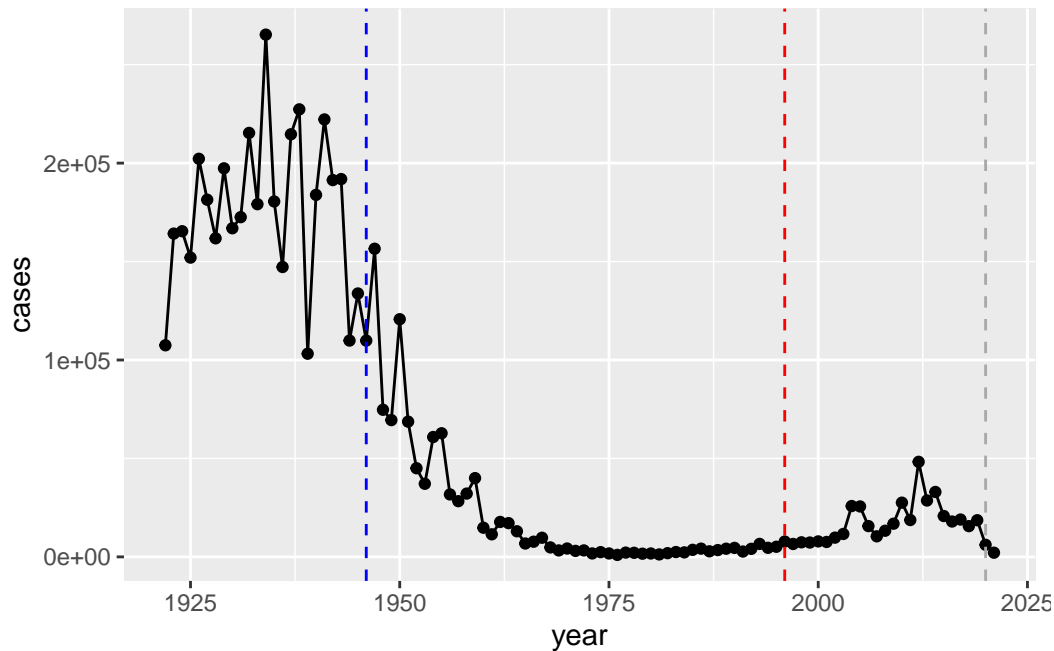
Q3. Describe what happened after the introduction of the aP vaccine? Do you have a possible explanation for the observed trend?

Cases were low then jumped up in 2004. This could've been caused by different things like the anti-vax movement or a shorter period of immunity when given the aP vs wP vaccine. However none of that is clear so more study needs to be done.

Exploring CMI-PB Data

Why is this vaccine-preventable disease on the upswing? To answer this question we need to investigate the mechanisms underlying waning protection against pertussis. This requires evaluation of pertussis-specific immune responses over time in wP and aP vaccinated individuals.

This is the goal of the CMI-PB project: https://www.cmi-pb.org/

It makes its data available via "API-endpoint" that return JSON format. We will use the 'jsonlite' package to access this data.

```
library(jsonlite)

# Data tables
```

```r
subject <- read_json("http://cmi-pb.org/api/subject", simplifyVector = T)
specimen <- read_json("http://cmi-pb.org/api/specimen", simplifyVector = T)
titer <- read_json("http://cmi-pb.org/api/v4/plasma_ab_titer", simplifyVector = T)
```

```r
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
```

```r
head(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
```

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

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

```
aP wP
60 58
```

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

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

```
Female    Male
    79      39
```

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                                          21   11
Black or African American                       2    0
More Than One Race                              9    2
Native Hawaiian or Other Pacific Islander       1    1
Unknown or Not Reported                        11    4
White                                          35   20
```

The most represented category is white female. For research like this a more diverse study group would be preferrable.

Side-note: Working with Dates

Dates can be super annoying to work with - the format and order is different around the world. Using the lubridate package makes it a LOT easier to calculate dates:

```r
library(lubridate)
```

```
Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

```r
today()
```

```
[1] "2023-12-08"
```

```r
today() - mdy("10-7-2002")
```

```
Time difference of 7732 days
```

```r
today() - ymd("2002-10-7")
```

```
Time difference of 7732 days
```

```r
time_length( today() - ymd("1961-11-9"), "years")
```

```
[1] 62.07803
```

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?

```r
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
```

```
# average wP age
subject %>%
  filter(infancy_vac == "wP") %>%
  summarize(meanAge = mean(time_length(today() - ymd(year_of_birth), "years")))
```

```
  meanAge
1 36.32977
```

```
# average aP age
subject %>%
  filter(infancy_vac == "aP") %>%
  summarize(meanAge = mean(time_length(today() - ymd(year_of_birth), "years")))
```

```
  meanAge
1 26.03304
```

(iii). A ten-year difference between wP and aP vaccination ages seems pretty significant.

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

```
# Calculating age at time of boost
subject$age <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)

# Converting time length to years
subject$age_years <- time_length(subject$age, "years")

subject$age_years
```

```
 [1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481 35.84942 34.14921
 [9] 20.56400 34.56263 30.65845 34.56263 19.56194 23.61944 27.61944 29.56331
[17] 36.69815 19.65777 22.73511 35.65777 33.65914 31.65777 25.73580 24.70089
[25] 28.70089 33.73580 19.73443 34.73511 19.73443 28.73648 27.73443 19.81109
[33] 26.77344 33.81246 25.77413 19.81109 18.85010 19.81109 31.81109 22.81177
[41] 31.84942 19.84942 18.85010 18.85010 19.90691 18.85010 20.90897 19.04449
[49] 20.04381 19.90691 19.90691 19.00616 19.00616 20.04381 20.04381 20.07940
[57] 21.08145 20.07940 20.07940 20.07940 32.26557 25.90007 23.90144 25.90007
[65] 28.91992 42.92129 47.07461 47.07461 29.07324 21.07324 21.07324 28.15058
[73] 24.15058 24.15058 21.14990 21.14990 31.20876 26.20671 32.20808 27.20876
[81] 26.20671 21.20739 20.26557 22.26420 19.32375 21.32238 19.32375 19.32375
```

```
 [89] 22.41752 20.41889 21.41821 19.47707 23.47707 20.47639 21.47570 19.47707
 [97] 35.90965 28.73648 22.68309 20.83231 18.83368 18.83368 27.68241 32.68172
[105] 27.68241 25.68378 23.68241 26.73785 32.73648 24.73648 25.79603 25.79603
[113] 25.79603 31.79466 19.83299 21.91102 27.90965 24.06297
```

Q9. With the help of a faceted boxplot or histogram (see below), 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`.
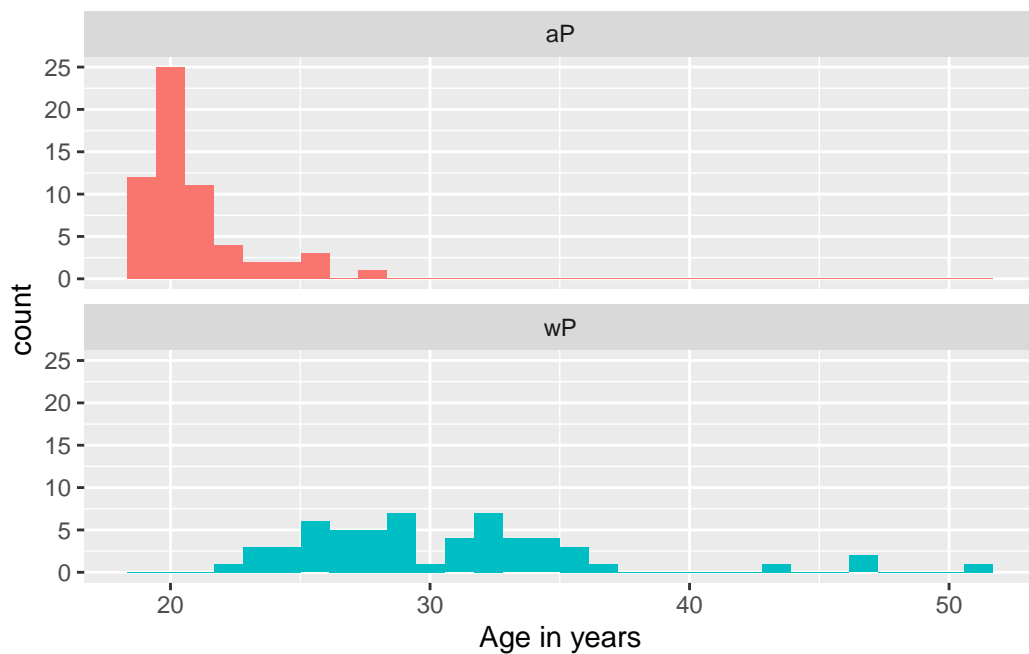
They are significantly different in age. aP subjects are all around 20 years old while none of the wP subjects are younger than 20.

Merge or Join Tables

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

```
head(meta)
```

```
  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 infancy_vac biological_sex
1                             0         Blood     1          wP         Female
2                             1         Blood     2          wP         Female
3                             3         Blood     3          wP         Female
4                             7         Blood     4          wP         Female
5                            14         Blood     5          wP         Female
6                            30         Blood     6          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
         age age_years
1 11212 days  30.69678
2 11212 days  30.69678
3 11212 days  30.69678
4 11212 days  30.69678
5 11212 days  30.69678
6 11212 days  30.69678
```

Antibody measurements are stored in the titer. However, it doesn't include any of the metadata:

```
head(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
```

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

```
dim(abdata)
```

```
[1] 41810    22
```

> 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 3240 7968 7968 7968 7968
```

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

Let's focus on one of these - IgG. We're gonna use dplyr to filter only this data.

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

```
  specimen_id isotype is_antigen_specific antigen       MFI MFI_normalised
1           1     IgG                TRUE      PT   68.56614       3.736992
2           1     IgG                TRUE     PRN  332.12718       2.602350
3           1     IgG                TRUE     FHA 1887.12263      34.050956
4          19     IgG                TRUE      PT   20.11607       1.096366
5          19     IgG                TRUE     PRN  976.67419       7.652635
6          19     IgG                TRUE     FHA   60.76626       1.096457
   unit lower_limit_of_detection subject_id actual_day_relative_to_boost
1 IU/ML                 0.530000          1                           -3
2 IU/ML                 6.205949          1                           -3
3 IU/ML                 4.679535          1                           -3
4 IU/ML                 0.530000          3                           -3
5 IU/ML                 6.205949          3                           -3
6 IU/ML                 4.679535          3                           -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             Unknown White    1983-01-01    2016-10-10 2020_dataset
5             Unknown White    1983-01-01    2016-10-10 2020_dataset
6             Unknown White    1983-01-01    2016-10-10 2020_dataset
        age age_years
1 11212 days  30.69678
2 11212 days  30.69678
3 11212 days  30.69678
4 12336 days  33.77413
5 12336 days  33.77413
```
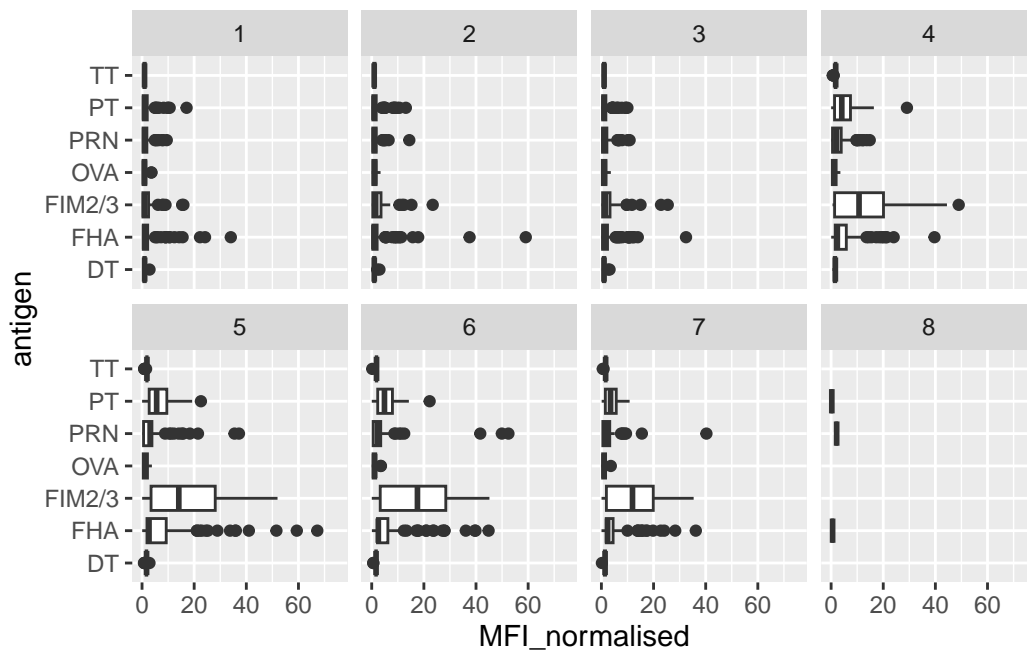
```
6 12336 days   33.77413
```

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)
```
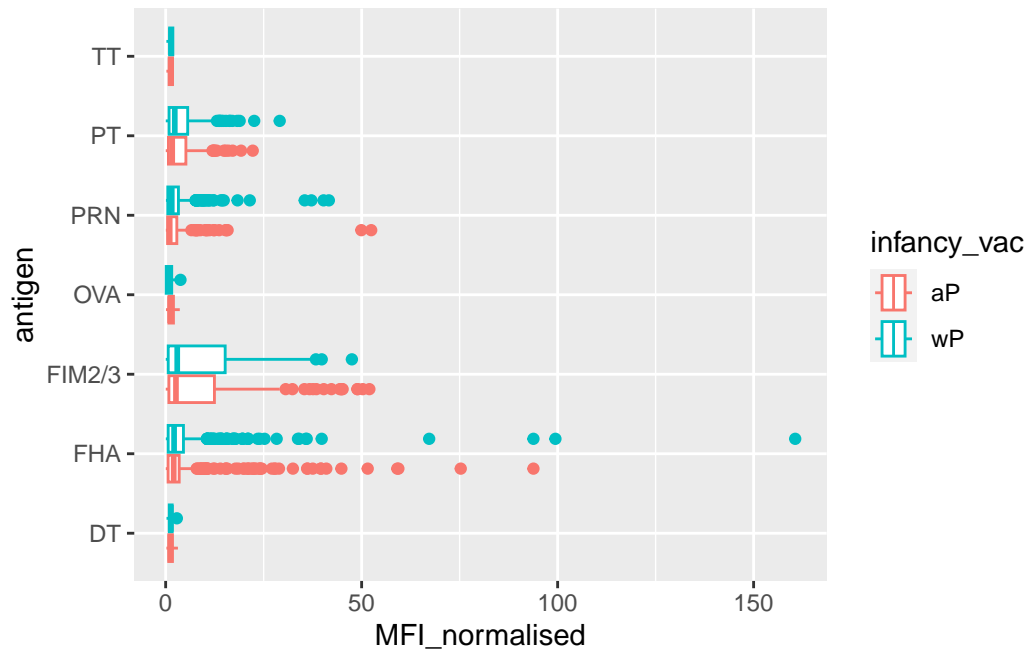
```
Warning: Removed 5 rows containing non-finite values (`stat_boxplot()`).
```


Other box plots we did in class:

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

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

Focus on IgG to the Pertussis Toxin (PT) antigen in the 2021 dataset:

```
igg.pt <- igg %>% filter(antigen == "PT", dataset =="2021_dataset")
```

```
ggplot(igg.pt) +
  aes(planned_day_relative_to_boost,
      MFI_normalised,
      col = infancy_vac,
      group = subject_id) +
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 0, linetype = "dashed") +
  geom_vline(xintercept = 14, linetype = "dashed")
```