Class 19

A07606689

```
library(datapasta)
```

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.

Pertussis is a bacterial infection that causes a severe cough. Often named "whooping cough"

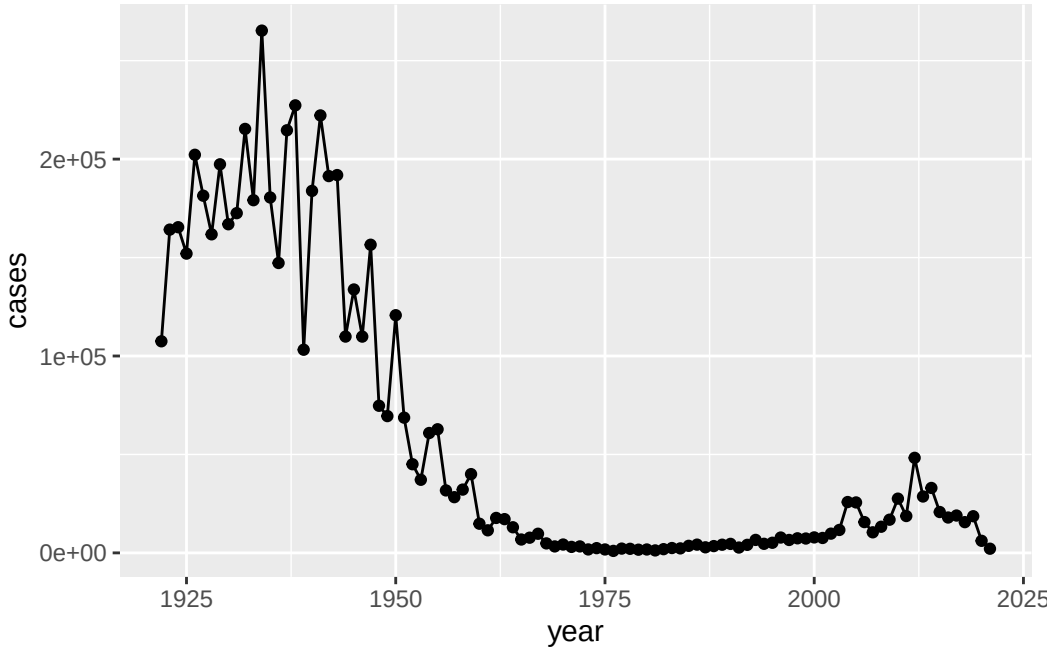
Lets have a look at case numbers of Pertussis in the US.

We can find these numbers on the [CDC website] (https://www.cdc.gov/pertussis/surv-reporting/cases-by-year.html)

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.4     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

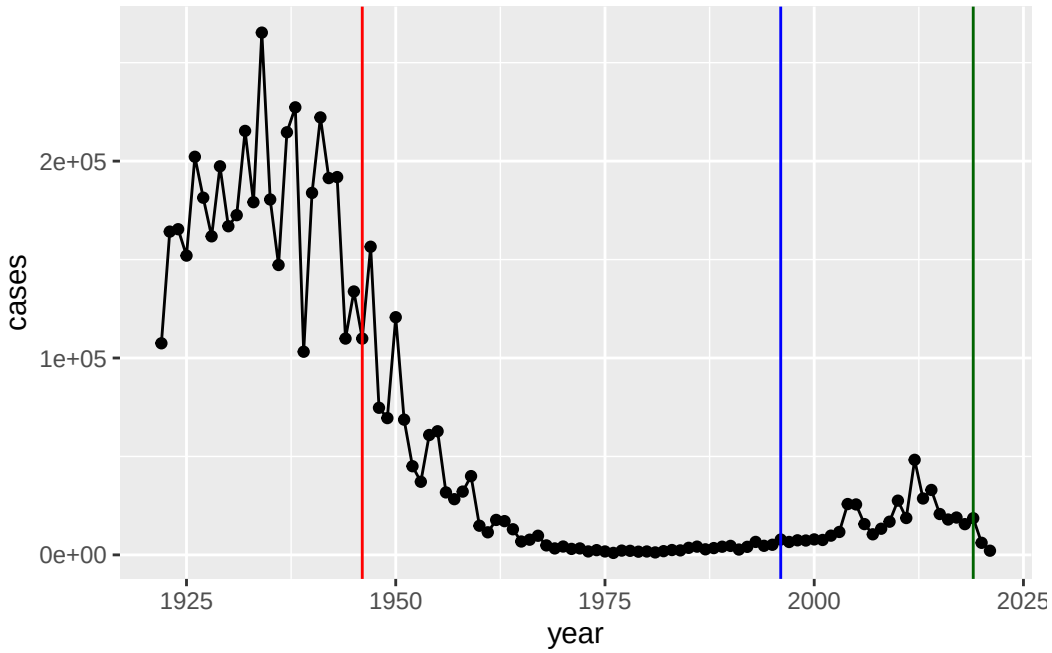
```
ggplot(cdc) +
  aes(x=year, y=cases) +
  geom_point() +
geom_line()
```



A tale of two vaccines (wP & aP)

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_point() +
geom_line() +
  geom_vline(xintercept =1946, color = "red") +
geom_vline(xintercept =1996, color = "blue") +
geom_vline(xintercept =2019, color = "darkgreen")
```



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

After the initial introduction of the aP vaccine there was a decrease of cases. However, the aP vaccine requires a booster (Tdap) every 10 years. The number of cases for Tdap started to increase after the 10 year period. It is expected that after 2019, the number of cases will shot up.

#CMI-PB project

The CMI-PB project collects and makes available data on the immune response to Pertussis booster vaccination.

We will access this via the API. We will use the *jsonlite* package to access the data using the 'read_json()' function.

```
library(jsonlite)
```

```
Attaching package: 'jsonlite'
```

```
The following object is masked from 'package:purrr':

    flatten
```

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

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

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

```
aP wP
60 58
```

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

```
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…)?

```
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

Side-Note: Working with dates

Q. Make a histogram of the subject age distribution and facet by infancy_vac

```
today() - mdy("09-12-1996")
```

```
Time difference of 9946 days
```

```
#"12-09-1996"

today() - dmy("13-01-1989")
```

```
Time difference of 12745 days
```

```
time_length( today() - ymd("1989-01-13"),  "years")
```

```
[1] 34.89391
```

```
subject$age <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
subject$age_years <- time_length(subject$age, "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? t.test

```
ap <- subject %>% filter(infancy_vac == "aP")

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

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     19      20      20      21      21      28
```

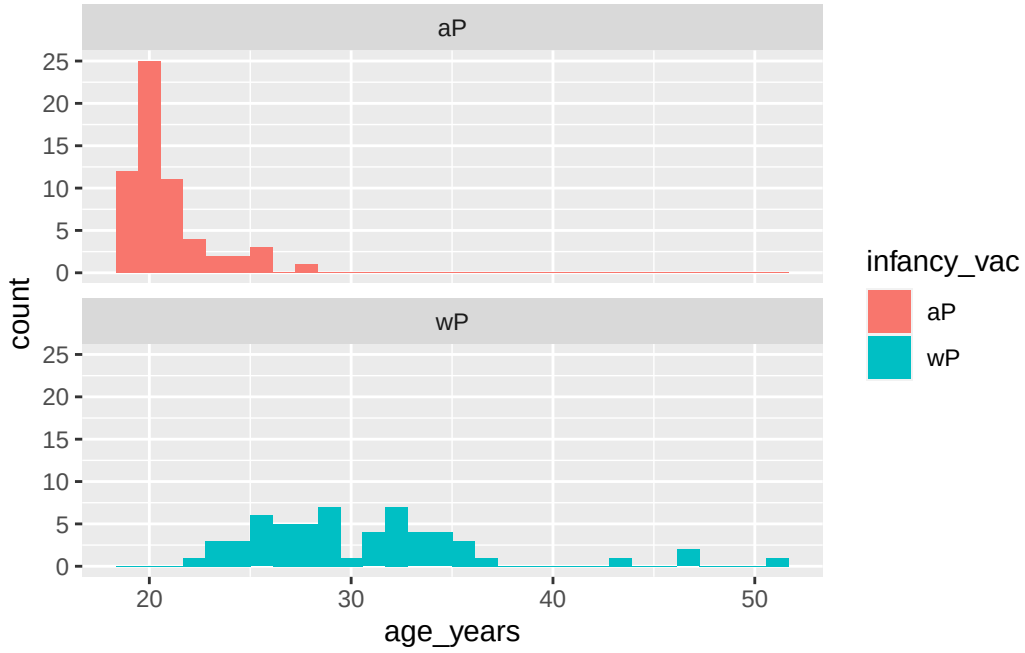
```
wp <- subject %>% filter(infancy_vac == "wP")

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

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   23      26      29      31      34      51
```

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

```
ggplot(subject) +
  aes(age_years,
    fill=infancy_vac) +
  facet_wrap(vars(infancy_vac), ncol = 1) +
  geom_histogram()
```

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



Q9. With the help of a faceted boxplot or histogram (see below), do you think these two groups are significantly different? There are 3 main datasets in the CMI-PB project at the time of writing:

```
table(subject$dataset)
```

```
2020_dataset 2021_dataset 2022_dataset
          60           36           22
```

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

We will have a wee peak at the tables

Joining multiple 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:

I want to merge (join) the specimen and subject tables together.

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

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

```
dim(meta)
```

```
[1] 939  15
```

```
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
```

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.

Now I want to join the merge (join) he titer and meta data

```
abdata <- inner_join(titer, meta)
```

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

```
dim(abdata)
```

```
[1] 41810    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
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
```

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
```

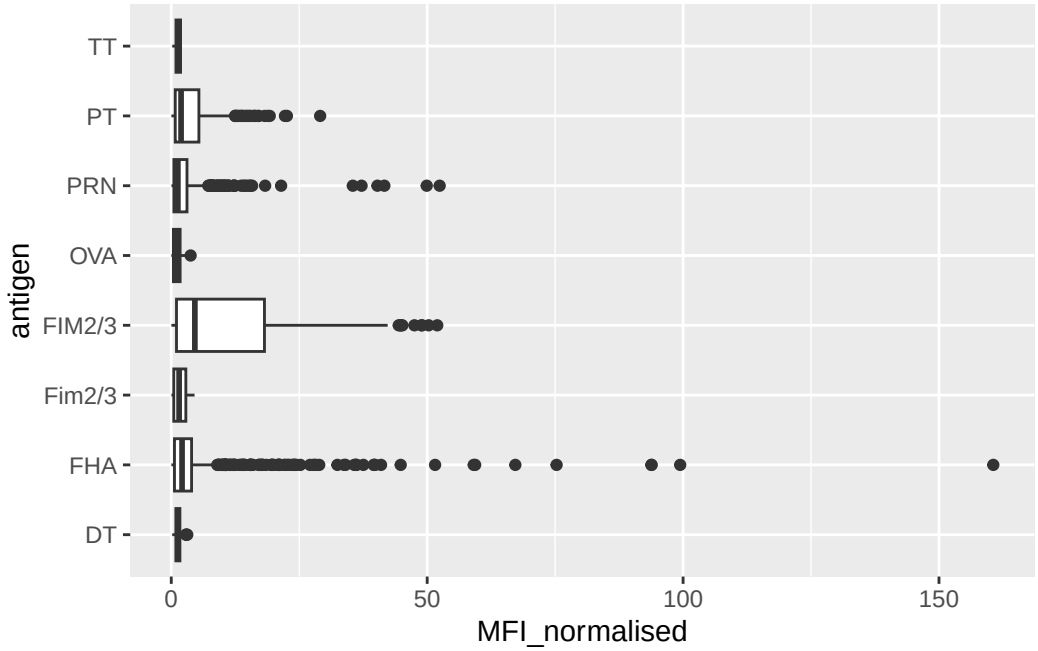
Q12. What are the different $dataset values in abdata and what do you notice about the number of rows for the most "recent" dataset? skipped in class

Examine IgG Ab titer levels

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

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()
```




```
oops <- abdata %>% filter(antigen=="Fim2/3")
table(oops$dataset)
```

```
2022_dataset
         315
```

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

```
2020_dataset 2021_dataset 2022_dataset
       31520         8085         2205
```

Select (or fliter) for the 2021 dataset and isotype IgG. I want time course ('day_relative_to_boost') of IgG levels ('MFI_normalised') for "PT" antigen.

```r
igpt.21 <- abdata %>% filter(dataset=="2021_dataset", isotype=="IgG", antigen=="PT")
ggplot(igpt.21) +
  aes(planned_day_relative_to_boost,
     MFI_normalised,
     col=infancy_vac) +
  geom_point() +
  geom_line(aes(group=subject_id), linewidth=0.5, alpha=0.5) +
  geom_smooth(se=FALSE, span=0.4, linewidth=3) +
  geom_vline(xintercept= 0) +
geom_vline(xintercept = 14)
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at -0.6
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 3.6
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 1.8382e-16
```

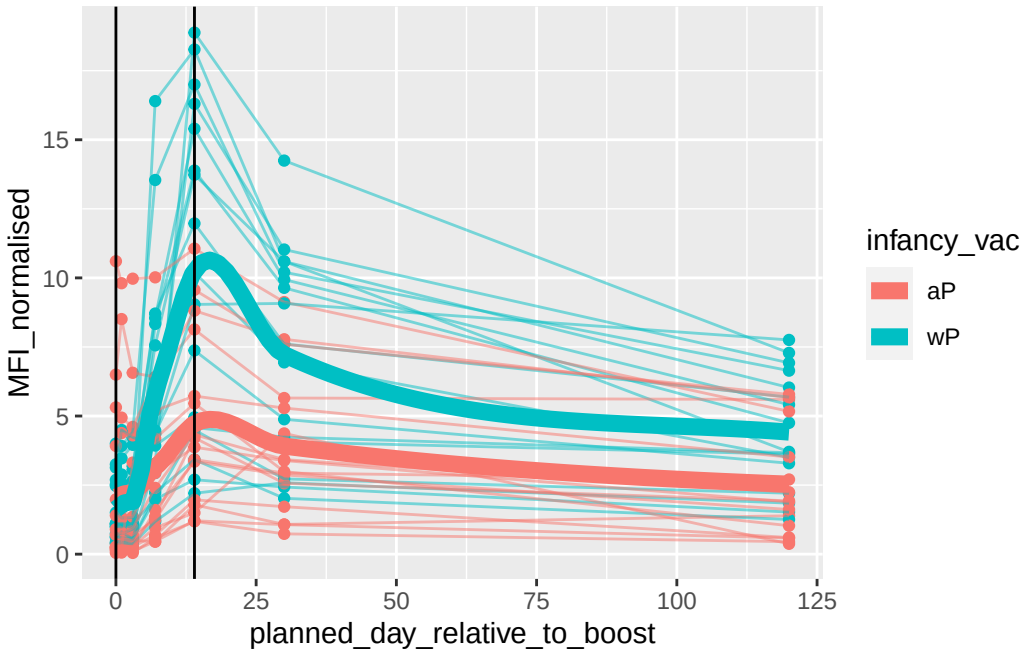
```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 11364
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at -0.6
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 3.6
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 1.4316e-16
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 11364
```




```
igpt.22 <- abdata %>% filter(dataset=="2022_dataset", isotype=="IgG", antigen=="PT")
ggplot(igpt.22) +
  aes(planned_day_relative_to_boost,
     MFI_normalised,
     col=infancy_vac) +
  geom_point() +
  geom_line(aes(group=subject_id), linewidth=0.5, alpha=0.5) +
  geom_smooth(se=FALSE, span=0.4, linewidth=3) +
  geom_vline(xintercept= 0) +
geom_vline(xintercept = 14)
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at -30.15
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 15.15
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 0
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 229.52

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at -30.15

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 15.15

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 0

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 229.52
```