Class 15 Pertussis SO COOL!

WBray A69034838

#This is an awesome topic! #First create a data frame using the CDC data set linked. #use datapasta to insert data frame from linked data here #So we want to web scrape this information from the website! #install datapasta from CRAN, then use addins (or tools -> addins) to select this function! #there's another package called rvst, but requires raw html and is therefore more effort to deal with

```
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, 2024L),
  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, 23544)
)
```

#installed Styler in otder to cleanup stuff! Didn’t use it though

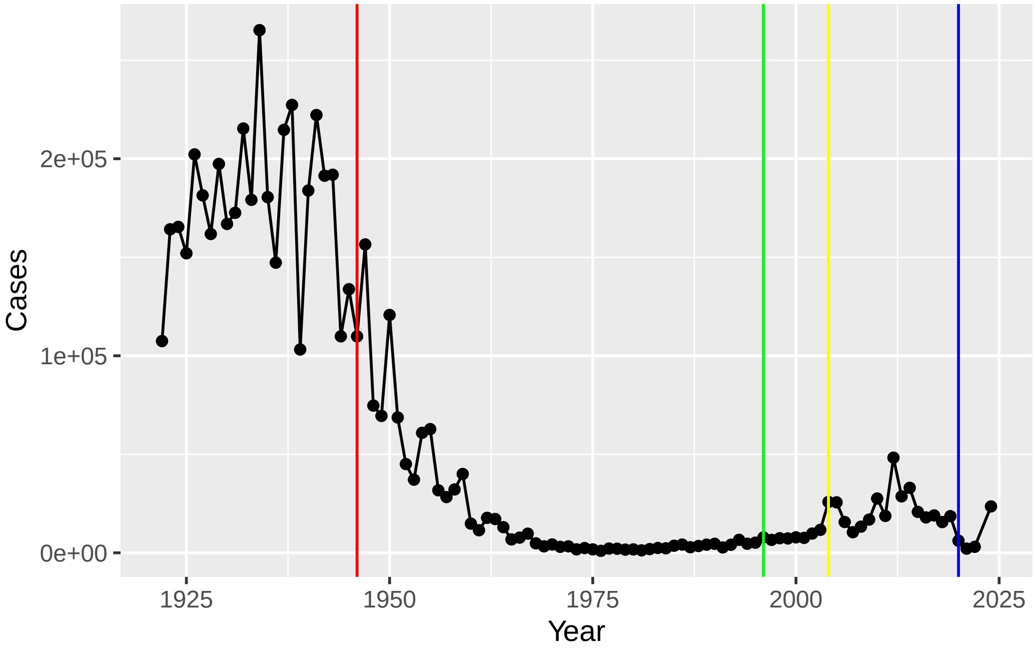
```
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.4.2
```

Question 1:

```
baseplot <- ggplot(cdc) + aes(x= Year, y= Cases) + geom_point() + geom_line()

baseplot + geom_vline(xintercept = 1946, col = "red") + geom_vline(xintercept = 1996, col = "
geom_vline(xintercept = 2020, col = "blue") +
geom_vline(xintercept = 2004, col = "yellow")
```



Question 2: The original vaccine was exceptionally effective; cases declined precipitously to almost nothing after introduction of the original vaccine. The aP vaccine may not be quite as effective, as we see some increase after its introduction, but this is also after a decade of anti-vaccine propaganda. Question 3: The aP vaccine may not possess the same duration of protection as the original whole cell killed vaccine.

#look like the acellular vaccine has attenuated long term efficacy, would not have showed up in clinical trials since the phoneomena only appeared a decade after rollout. Why? This is where Barry and company come in… #CMI-PB; can study individuals who had different types of vaccines to prime their immune response (boosting with aP vaccine) #making data available to the public, has challenges for the scientific community…is HLA haplotype listed? (in fact they are doing whole genome sequencing, and have PBMC transcriptomics!) #check understanidng the data section #will need to check the API in order to pull down salient data. # This project collects and makes available data abuot the immune responseto the Pertussis vaccine.. #Can be accessedvia API which returns JSON format (key:value pairs) #therefore install JSOnlite package

```
library(jsonlite)
```

```
Warning: package 'jsonlite' was built under R version 4.4.2
```

```r
subject <- read_json("http://cmi-pb.org/api/v5/subject", simplifyVector = TRUE)
```

```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
```

Question 4: how many subjects are in this dataset? - 172.

```r
nrow(subject)
```

```
[1] 172
```

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

```

Female   Male 
   112     60 
```

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

```

aP wP 
87 85 
```

Question 5: 60 male, 112 female

#remember, table can do moultiple variables at one, but they are separated inside of the parentheticals by a comma, NOT nested

```
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
```

Question 6: - definitely does NOT reflect the US population overall; skewed toward UCSD students that needed the money and were willing to go into a hospital during the pandemic.

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

```

2020_dataset 2021_dataset 2022_dataset 2023_dataset 
          60           36           22           54 
```

#read in more data!

```
specimen <- read_json("http://cmi-pb.org/api/v5/specimen", simplifyVector = TRUE)
ab_titer <- read_json("http://cmi-pb.org/api/v5/plasma_ab_titer", simplifyVector = TRUE)
PBMC <- read_json("http://cmi-pb.org/api/v5/pbmc_gene_expression?limit=25", simplifyVector =
```

#let’s check the head of these to see what commonalities we can find..

```
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
```

BARRY SKIPPED QUESTIONS 7/8

#practice some dplyr; let's combine these various table with the join command, we want antibody measurements combined with subject Id! Super cool, very important! Question 9:

```
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)`
```

Question 10:

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

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

```
nrow(abdata)
```

```
[1] 52576
```

#want to make plots with the various variables

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

#we want to see how many different isotypes etc. in this file Question 11:

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

```

  IgE   IgG  IgG1  IgG2  IgG3  IgG4 
 6698  5389 10117 10124 10124 10124 
```

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

```

2020_dataset 2021_dataset 2022_dataset 2023_dataset 
       31520         8085         7301         5670 
```

Question 12: Values decline over time; not getting as many follow-up appointments as they would like!

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

```

    ACT   BETV1      DT   FELD1     FHA  FIM2/3   LOLP1     LOS Measles     OVA 
```

```
   1970    1970    4978    1970    5372    4978    1970    1970    1970    4978
    PD1     PRN      PT     PTM   Total      TT
   1970    5372    5372    1970     788    4978
```

#TT is tetanus toxoid, pertussis toxin is PT, FIM2/3 is filamentous hemaglutinin; don't want to see spikes in measles (ctrl) #Let's begin our filtration wit IgG

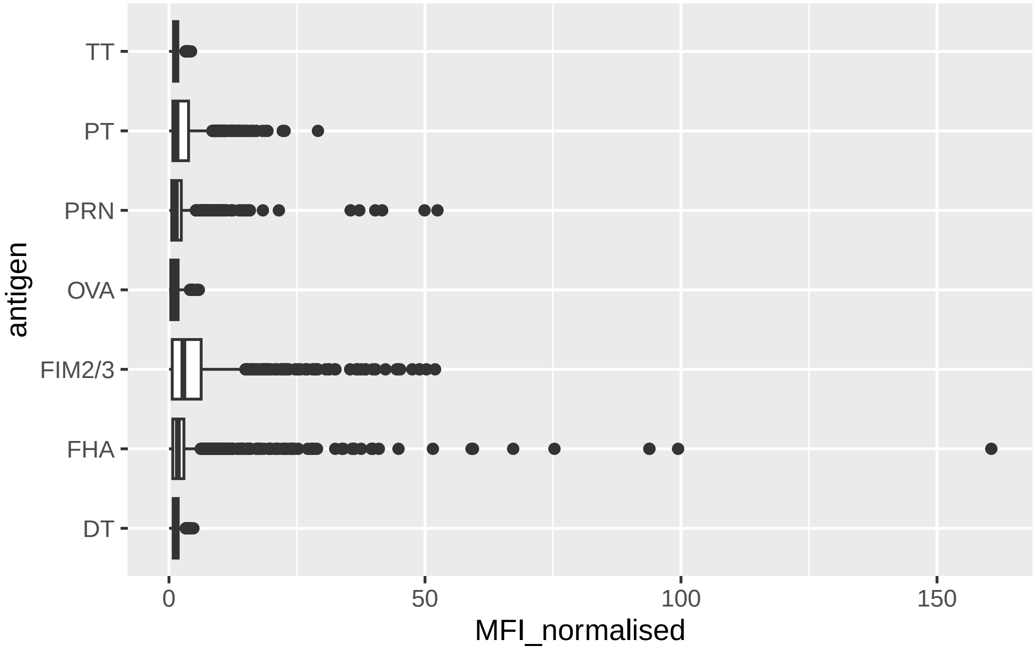
```
igg <- filter(abdata, isotype == "IgG")
head(igg)
```

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

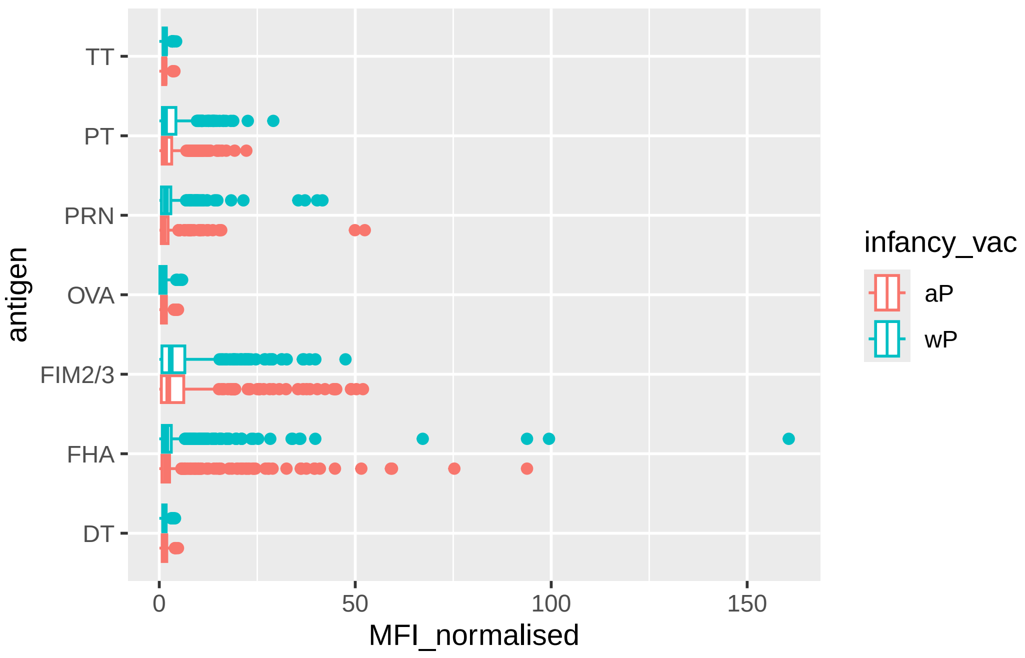
```
4                  0.530000
5                  6.205949
6                  4.679535
```

Question 13:

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



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



#we'd really like to see the time dependency in this case; specifically in relation to booster administration

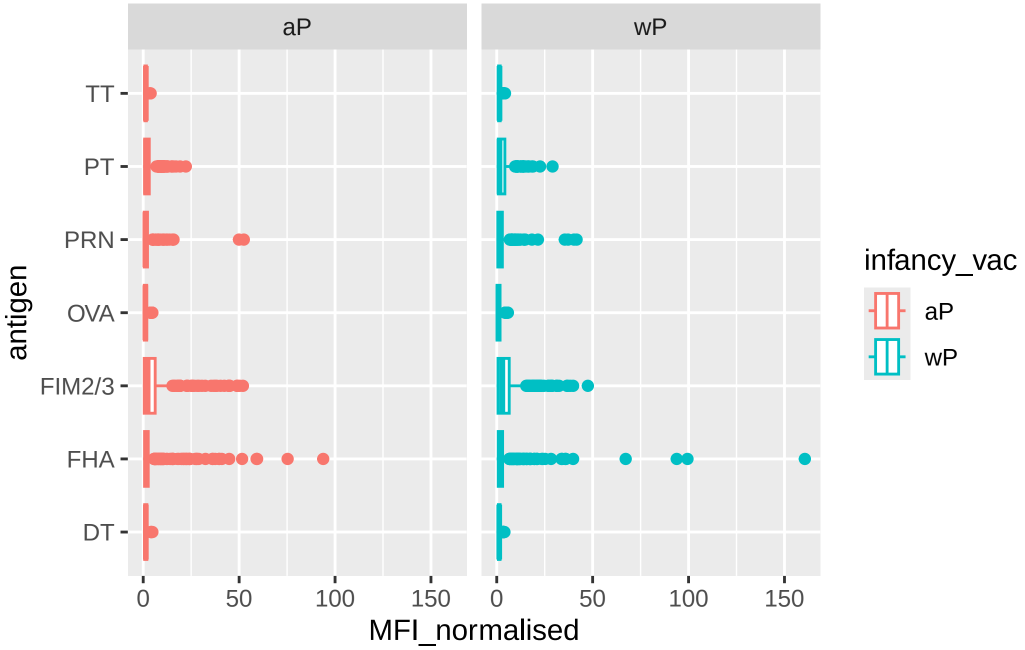
```
table(abdata$visit)
```

```

   1    2    3    4    5    6    7    8    9   10   11   12 
8280 8280 8420 6565 6565 6210 5810  815  735  686  105  105 
```

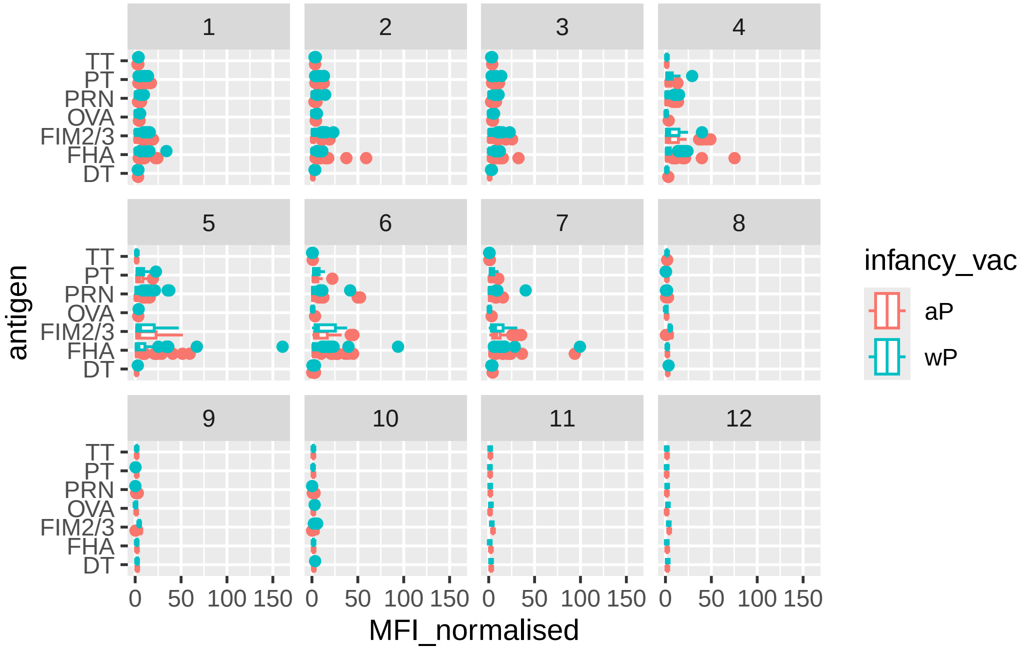
#try to utilize the facet wrap with the infancy data..

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



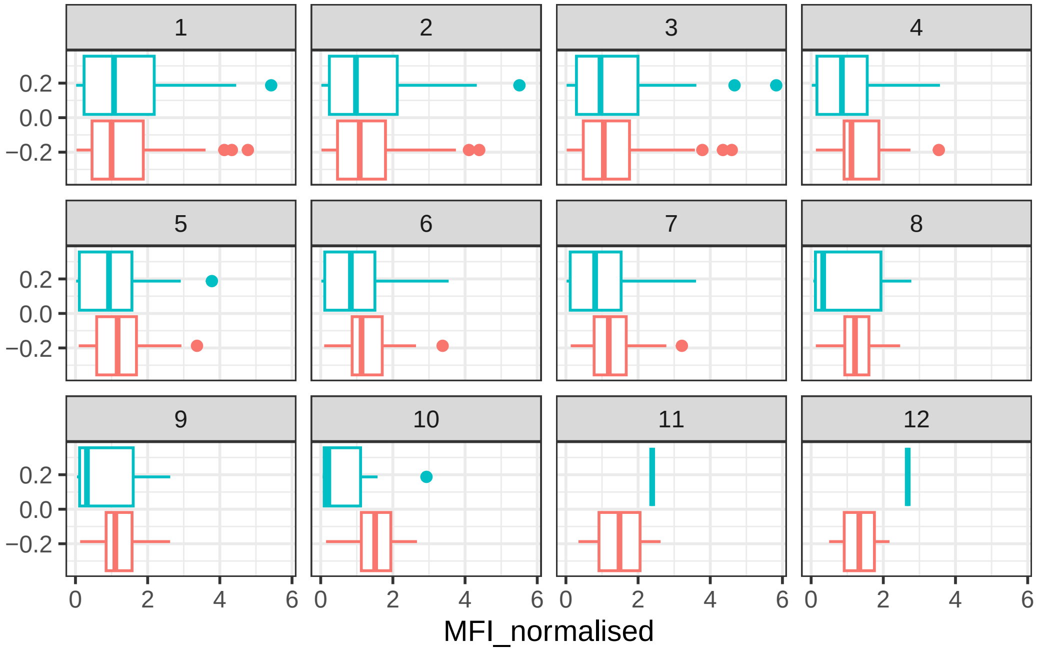
#Nice; ideally would be larger with transformed/scaled axes, but that's for another time. Question 14: Responses decline with time, with the individuals responding positively to antigens in the booster!

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

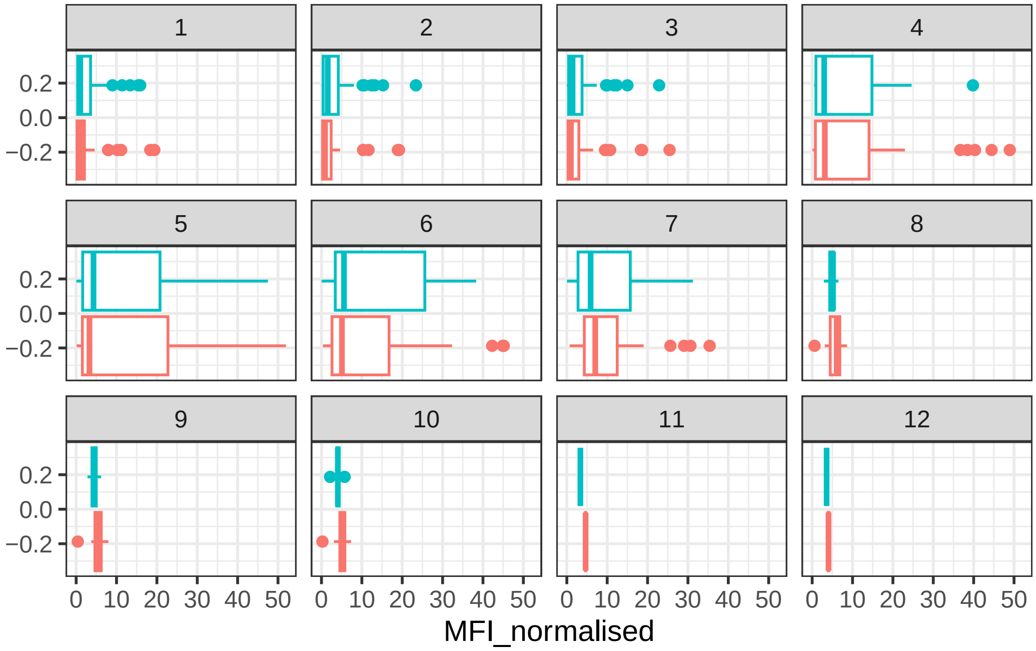


Question 15:

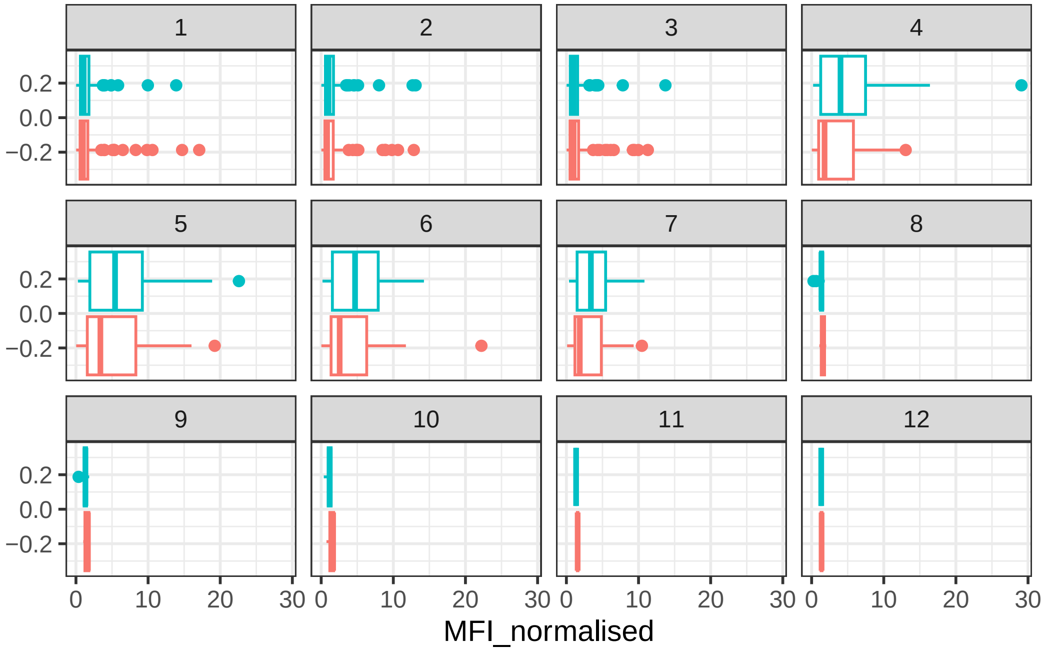
```
filter(igg, antigen=="OVA") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



```
filter(igg, antigen=="FIM2/3") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



```
filter(igg, antigen=="PT") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



Q16: The PT response clearly increases post booster, but then declines once again to baseline. Ova remains rather elevated (continuous exposure, hence positive control antigen).

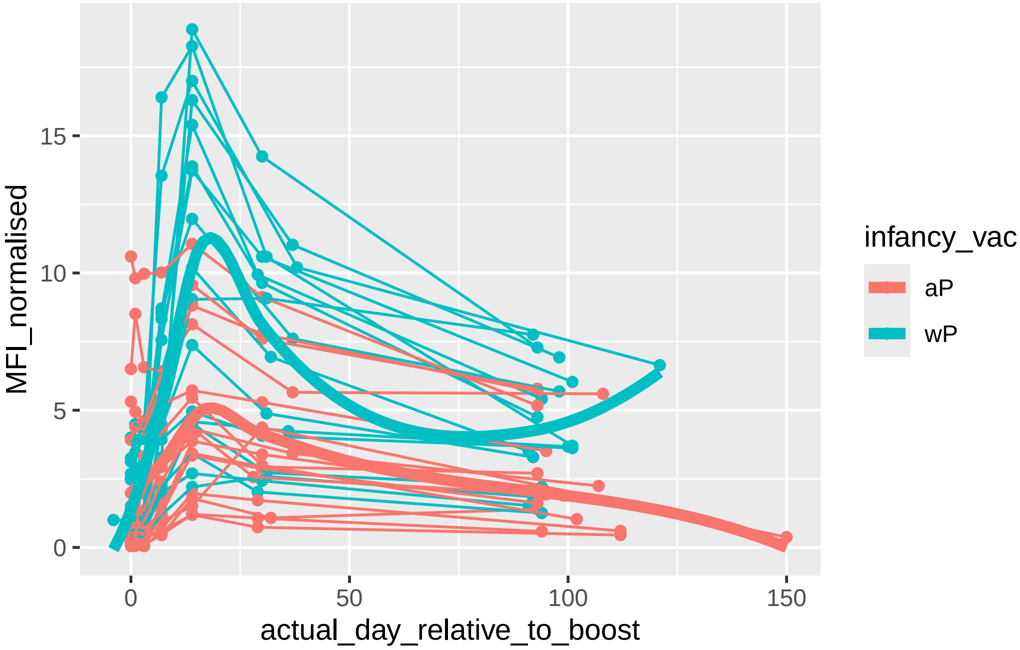
```
anti <- filter(igg, antigen == "PT", dataset == "2021_dataset")
head(anti)
```

```
  subject_id infancy_vac biological_sex              ethnicity
1         61          wP         Female Not Hispanic or Latino
2         61          wP         Female Not Hispanic or Latino
3         61          wP         Female Not Hispanic or Latino
4         61          wP         Female Not Hispanic or Latino
5         61          wP         Female Not Hispanic or Latino
6         61          wP         Female Not Hispanic or Latino
                     race year_of_birth date_of_boost      dataset specimen_id
1 Unknown or Not Reported    1987-01-01    2019-04-08 2021_dataset         468
2 Unknown or Not Reported    1987-01-01    2019-04-08 2021_dataset         469
3 Unknown or Not Reported    1987-01-01    2019-04-08 2021_dataset         470
4 Unknown or Not Reported    1987-01-01    2019-04-08 2021_dataset         471
5 Unknown or Not Reported    1987-01-01    2019-04-08 2021_dataset         472
6 Unknown or Not Reported    1987-01-01    2019-04-08 2021_dataset         473
  actual_day_relative_to_boost planned_day_relative_to_boost specimen_type
1                           -4                             0         Blood
2                            1                             1         Blood
```

```
3                               3                                 3         Blood
4                               7                                 7         Blood
5                              14                                14         Blood
6                              30                                30         Blood
  visit isotype is_antigen_specific antigen    MFI MFI_normalised unit
1     1     IgG               FALSE      PT 112.75      1.0000000  MFI
2     2     IgG               FALSE      PT 111.25      0.9866962  MFI
3     3     IgG               FALSE      PT 125.50      1.1130820  MFI
4     4     IgG               FALSE      PT 224.25      1.9889135  MFI
5     5     IgG               FALSE      PT 304.00      2.6962306  MFI
6     6     IgG               FALSE      PT 274.00      2.4301552  MFI
  lower_limit_of_detection
1                 5.197441
2                 5.197441
3                 5.197441
4                 5.197441
5                 5.197441
6                 5.197441
```

```
ggplot(anti) + aes(actual_day_relative_to_boost, MFI_normalised, col=infancy_vac, group = su
```

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



Question 17: Quite an interesting trend. There does appear to be a difference between acellular and whole Pertussis vaccines, as the latter seems to correlate with higher titers in response to the booster! #submit as pdf