

Class18: Pertussis

Ani A16647613

We will begin by examining the cases of Pertussis tracked by the CDC: <https://www.cdc.gov/pertussis/surv-reporting/cases-by-year.html>

In order to input the table from the yearly cases of Pertussis, we can use the datapasta package to add the data into R:

Q1

```
library(datapasta)
```

Warning: package 'datapasta' was built under R version 4.3.3

```
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),
No..Reported.Pertussis.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)
)

```

```
head(cdc)
```

| | Year | No..Reported.Pertussis.Cases |
|---|------|------------------------------|
| 1 | 1922 | 107473 |
| 2 | 1923 | 164191 |
| 3 | 1924 | 165418 |
| 4 | 1925 | 152003 |
| 5 | 1926 | 202210 |
| 6 | 1927 | 181411 |

Q2 In order to make a plot of cases per year:

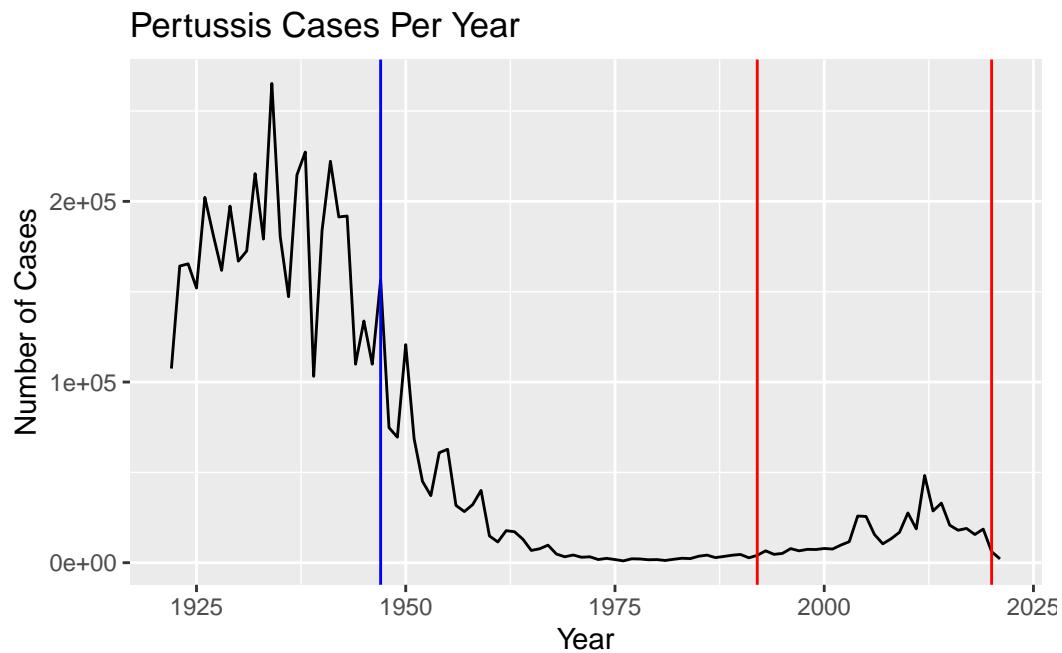
```

library(ggplot2)

ggplot(cdc) +
  aes(Year, No..Reported.Pertussis.Cases) +
  geom_line() +
  geom_vline(xintercept=1947, col="blue") +
  geom_vline(xintercept=1992, col="red") +

```

```
geom_vline(xintercept=2020, col="red") +
labs(x= "Year", y= "Number of Cases", title="Pertussis Cases Per Year")
```



Q3 After the aP vaccine, we notice an increase in Pertussis cases in the following years

Access Data from CMI-PB project

The database uses an API to return JSON format data.

In order to analyze the data we will install the `jsonlite` package

```
library(jsonlite)
```

Warning: package 'jsonlite' was built under R version 4.3.3

```
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 wP (older whole-cell vaccine) individuals versus aP (newer acellular vaccine) individuals are in this data:

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

```
aP wP
60 58
```

Q5

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

```
Female    Male
    79     39
```

Q6

```
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 |

what is the number of Individuals by biological sex and race? shown above! Overall shows lack of accuracy in the sample demographic

Working with Dates

Using the lubridate package we can ease the pain of working with dates

```
library(lubridate)
```

Warning: package 'lubridate' was built under R version 4.3.3

Attaching package: 'lubridate'

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

```
date, intersect, setdiff, union
```

```
today()
```

```
[1] "2024-03-07"
```

How many days passed since new year 2000

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

Time difference of 8832 days

Based on my birthday:

```
today() - ymd("2002-09-24")
```

Time difference of 7835 days

Q8 Determining age of all individuals at boost time:

```
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
```

To find the age of everyone in our dataset:

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

subject$age
```

```
[1] 38.17933 56.18070 41.18001 36.18070 33.18001 36.18070 43.17864 39.17864
[9] 28.18070 42.17933 38.17933 42.17933 27.17864 31.17864 35.17864 37.18001
[17] 44.18070 27.17864 30.17933 43.17864 41.18001 39.17864 33.18001 32.18070
[25] 36.18070 41.18001 27.17864 42.17933 27.17864 36.18070 35.17864 27.17864
[33] 34.17933 41.18001 33.18001 27.17864 26.17933 27.17864 39.17864 30.17933
[41] 39.17864 27.17864 26.17933 26.17933 27.17864 26.17933 28.18070 26.17933
[49] 27.17864 27.17864 27.17864 26.17933 26.17933 27.17864 27.17864 27.17864
[57] 28.18070 27.17864 27.17864 27.17864 37.18001 31.17864 29.18001 31.17864
[65] 34.17933 48.18070 52.18070 52.18070 34.17933 26.17933 26.17933 33.18001
[73] 29.18001 29.18001 26.17933 26.17933 36.18070 31.17864 37.18001 32.18070
[81] 31.17864 26.17933 25.18001 27.17864 24.18070 26.17933 24.18070 24.18070
[89] 27.17864 25.18001 26.17933 24.18070 28.18070 25.18001 26.17933 24.18070
[97] 38.17933 31.17864 25.18001 23.17864 21.18001 21.18001 30.17933 35.17864
[105] 30.17933 28.18070 26.17933 29.18001 35.17864 27.17864 28.18070 28.18070
[113] 28.18070 34.17933 22.17933 24.18070 30.17933 26.17933
```

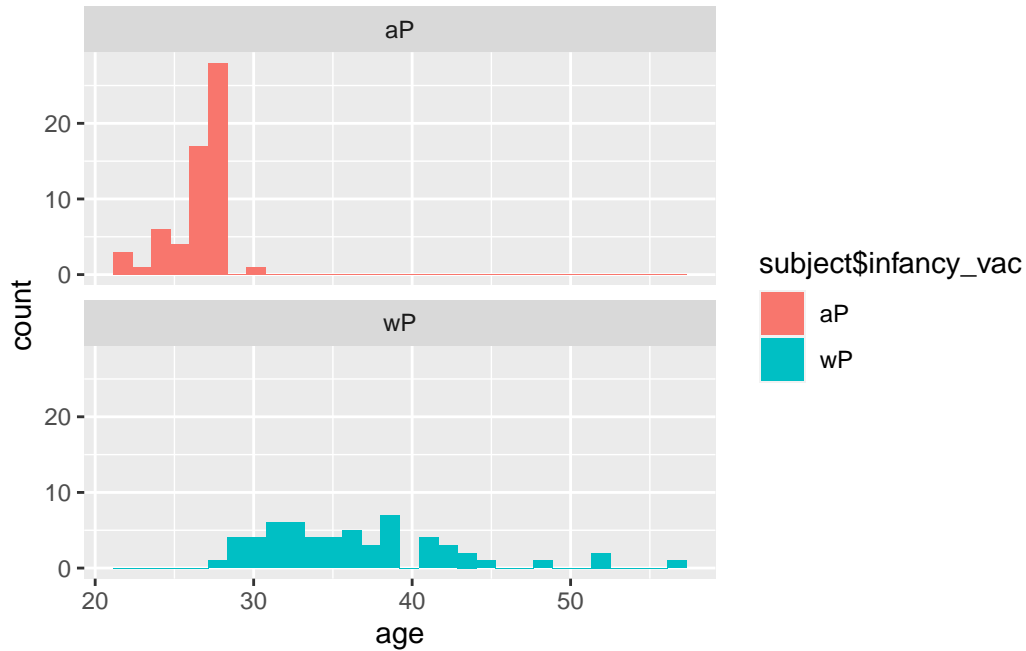
Q9

```
ggplot(subject) +
  aes(age, fill=subject$infancy_vac) +
  geom_histogram() +
```

```
facet_wrap(vars(subject$infancy_vac), nrow=2)
```

Warning: Use of `subject\$infancy_vac` is discouraged.
i Use `infancy_vac` instead.

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



Get more Data from CMI-PB

```
specimen <- read_json("http://cmi-pb.org/api/specimen", 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 |

Let's **join** both *subject* and *specimen* tables to make a “meta” table with our metadata by using the dplyr join function.

```
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 | 38.17933 | 1 |
| 2 | 1986-01-01 | 2016-09-12 | 2020_dataset | 38.17933 | 2 |
| 3 | 1986-01-01 | 2016-09-12 | 2020_dataset | 38.17933 | 3 |
| 4 | 1986-01-01 | 2016-09-12 | 2020_dataset | 38.17933 | 4 |
| 5 | 1986-01-01 | 2016-09-12 | 2020_dataset | 38.17933 | 5 |
| 6 | 1986-01-01 | 2016-09-12 | 2020_dataset | 38.17933 | 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 |

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

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 |

Q10 One more `inner_join()` to add all our metadata from `meta` and `ab_titer`

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

Joining with `by = join_by(specimen_id)`

```
dim(abdata)
```

```
[1] 41775    21
```

```
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 | infancy_vac | biological_sex |
|---|-------|--------------------------|------------|-------------|----------------|
| 1 | UG/ML | 2.096133 | 1 | wP | Female |
| 2 | IU/ML | 29.170000 | 1 | wP | Female |
| 3 | IU/ML | 0.530000 | 1 | wP | Female |
| 4 | IU/ML | 6.205949 | 1 | wP | Female |
| 5 | IU/ML | 4.679535 | 1 | wP | Female |
| 6 | IU/ML | 2.816431 | 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 | actual_day_relative_to_boost | planned_day_relative_to_boost |
|---|----------|------------------------------|-------------------------------|
| 1 | 38.17933 | -3 | 0 |
| 2 | 38.17933 | -3 | 0 |
| 3 | 38.17933 | -3 | 0 |
| 4 | 38.17933 | -3 | 0 |
| 5 | 38.17933 | -3 | 0 |
| 6 | 38.17933 | -3 | 0 |

| | specimen_type | visit |
|--|---------------|-------|
|--|---------------|-------|

| | | |
|---|-------|---|
| 1 | Blood | 1 |
| 2 | Blood | 1 |
| 3 | Blood | 1 |
| 4 | Blood | 1 |
| 5 | Blood | 1 |
| 6 | Blood | 1 |

Q11

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

| IgE | IgG | IgG1 | IgG2 | IgG3 | IgG4 |
|------|------|------|------|------|------|
| 6698 | 3233 | 7961 | 7961 | 7961 | 7961 |

Q12

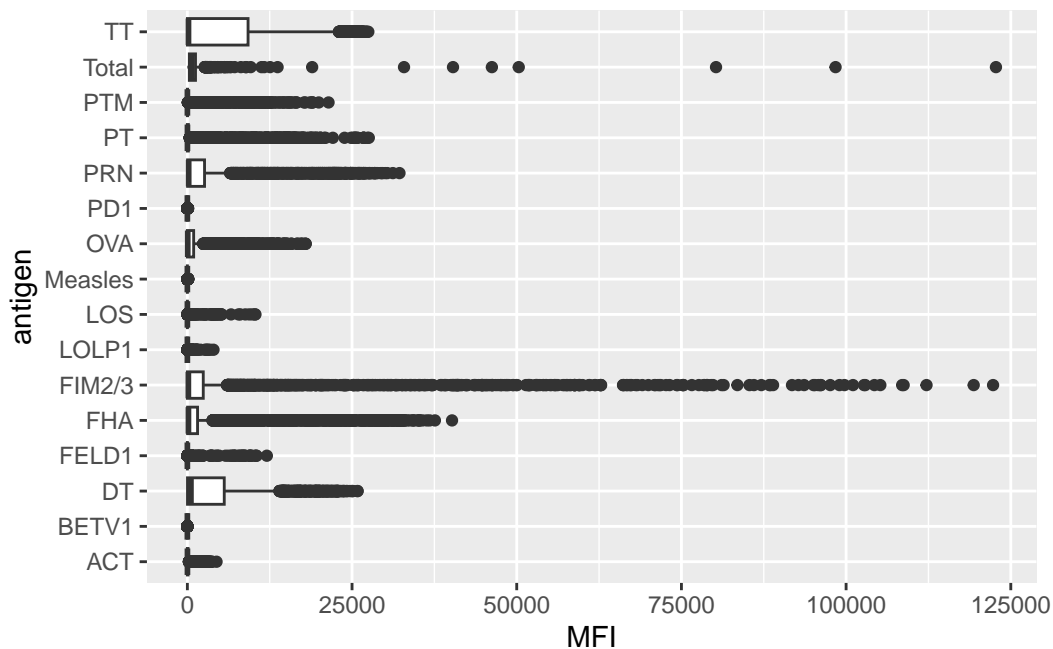
```
table(abdata$dataset)
```

| 2020_dataset | 2021_dataset | 2022_dataset |
|--------------|--------------|--------------|
| 31520 | 8085 | 2170 |

Q13

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

Warning: Removed 1 rows containing non-finite values (`stat_boxplot()`).



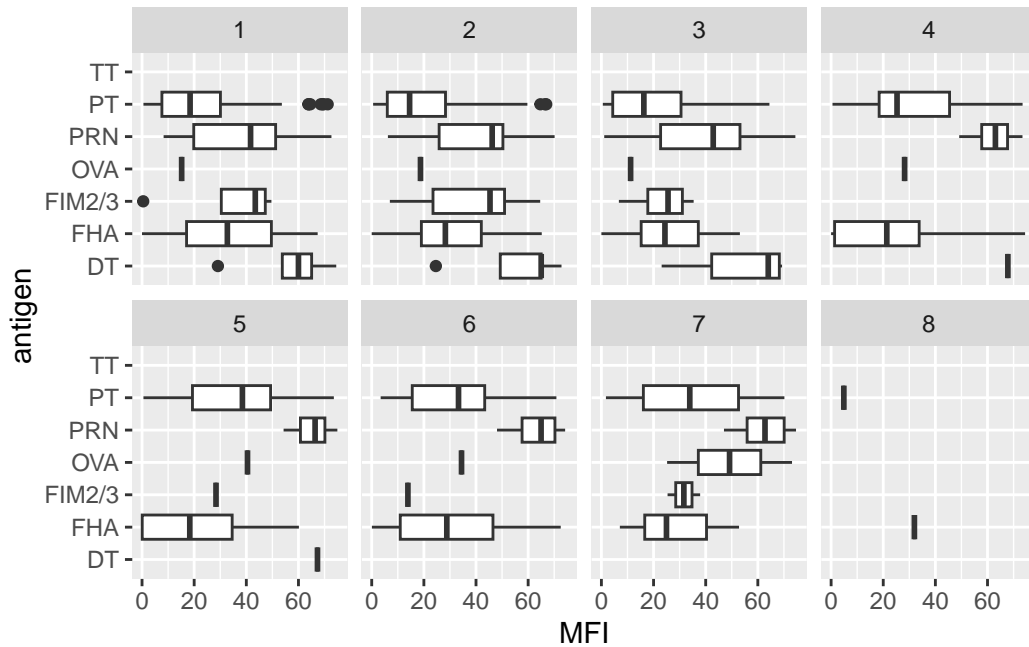
```
igg <- abdata %>% filter(isotype == "IgG")
head(igg)
```

| 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 | infancy_vac | biological_sex | | |
| 1 IU/ML | 0.530000 | 1 | wP | Female | | |
| 2 IU/ML | 6.205949 | 1 | wP | Female | | |
| 3 IU/ML | 4.679535 | 1 | wP | Female | | |
| 4 IU/ML | 0.530000 | 3 | wP | Female | | |
| 5 IU/ML | 6.205949 | 3 | wP | Female | | |
| 6 IU/ML | 4.679535 | 3 | 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 | actual_day_relative_to_boost | planned_day_relative_to_boost | | |
| 1 | 38.17933 | | -3 | | 0 |
| 2 | 38.17933 | | -3 | | 0 |
| 3 | 38.17933 | | -3 | | 0 |
| 4 | 41.18001 | | -3 | | 0 |
| 5 | 41.18001 | | -3 | | 0 |
| 6 | 41.18001 | | -3 | | 0 |
| | specimen_type | visit | | | |
| 1 | Blood | 1 | | | |
| 2 | Blood | 1 | | | |
| 3 | Blood | 1 | | | |
| 4 | Blood | 1 | | | |
| 5 | Blood | 1 | | | |
| 6 | Blood | 1 | | | |

```
ggplot(igg) +
  aes(MFI, antigen) +
  geom_boxplot() +
  xlim(0,75) +
  facet_wrap(vars(visit), nrow=2)
```

Warning: Removed 2514 rows containing non-finite values (`stat_boxplot()`).



Q14 The antigens that show different levels of IgG antibody titers are FHA, FIM2/3, and PRN for the most part.

Our first exploratory plot:

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

| | | | | | | | | | |
|------|-------|------|-------|-------|--------|-------|------|---------|------|
| ACT | BETV1 | DT | FELD1 | FHA | FIM2/3 | LOLP1 | LOS | Measles | OVA |
| 1970 | 1970 | 3435 | 1970 | 3829 | 3435 | 1970 | 1970 | 1970 | 3435 |
| PD1 | PRN | PT | PTM | Total | TT | | | | |
| 1970 | 3829 | 3829 | 1970 | 788 | 3435 | | | | |

Why are certain antigens and not others very variable in their detected levels here?

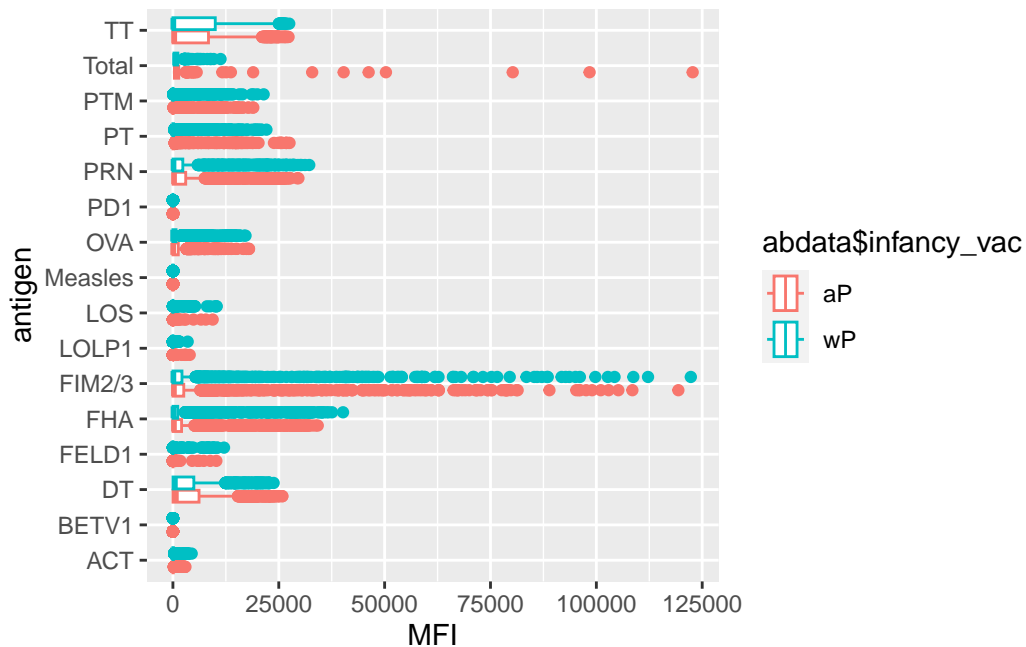
The aP vaccine contains the components of the cell that has high levels of antigen response

Can you facet or even just color by infancy_vac? Is there some difference?

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

Warning: Use of `abdata\$infancy_vac` is discouraged.
 i Use `infancy_vac` instead.

Warning: Removed 1 rows containing non-finite values (`stat_boxplot()`).



There are apparent differences between aP and wP but in general it is hard to tell with this whole dataset overview.

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

```
2020_dataset 2021_dataset 2022_dataset
      31520      8085      2170
```

Now to focus in on just the 2021_dataset

```
abdata.21 <- filter(abdata, dataset == "2021_dataset")
table(abdata.21$dataset)
```

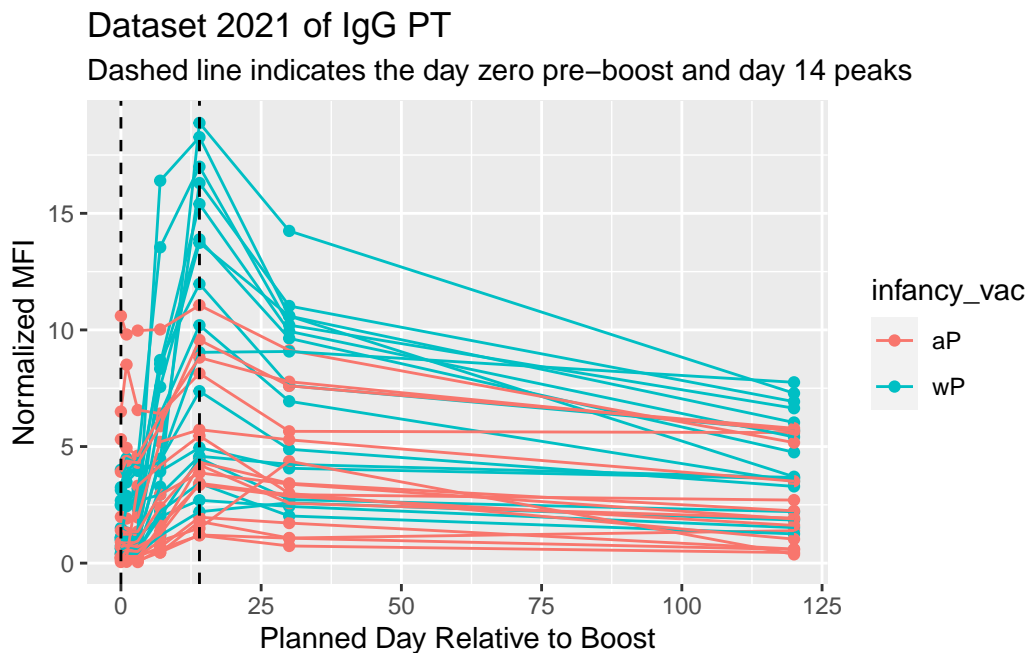
2021_dataset
8085

Focusing in on PT antigen IgG levels

```
pt.21 <- filter(abdata.21, isotype == "IgG", antigen == "PT")
```

now for plotting days (time) relative to normalized MFI

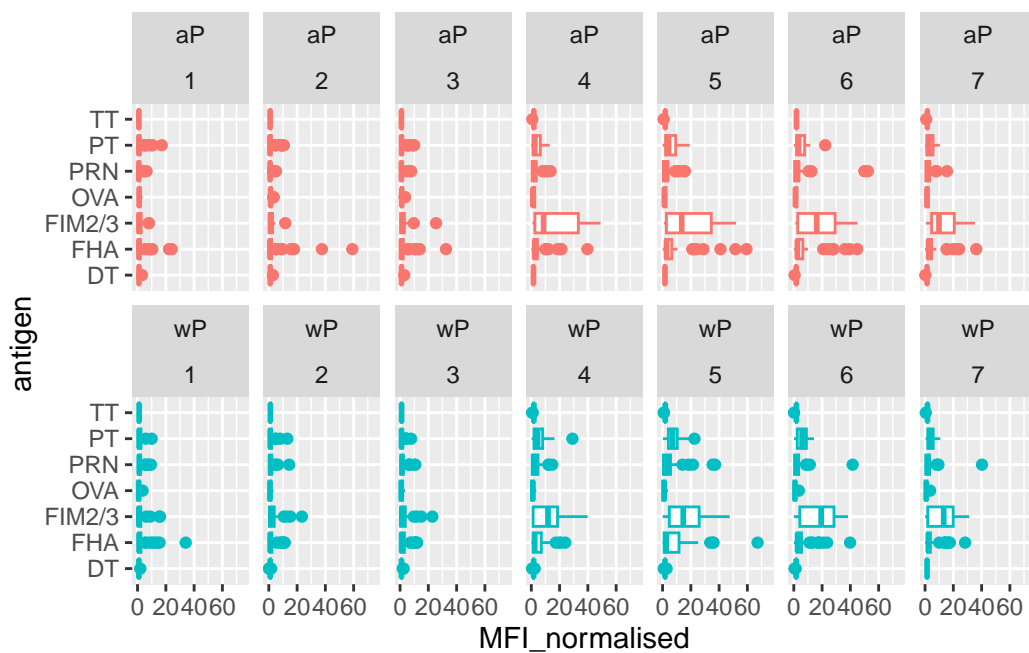
```
ggplot(pt.21) +  
  aes(x=planned_day_relative_to_boost,  
      y=MFI_normalised,  
      col=infancy_vac,  
      group= subject_id) +  
  geom_point() +  
  geom_line() +  
  geom_vline(xintercept=0, linetype="dashed") +  
  geom_vline(xintercept=14, linetype="dashed") +  
  labs(x= "Planned Day Relative to Boost", y= "Normalized MFI",  
       title= "Dataset 2021 of IgG PT",  
       subtitle="Dashed line indicates the day zero pre-boost and day 14 peaks")
```



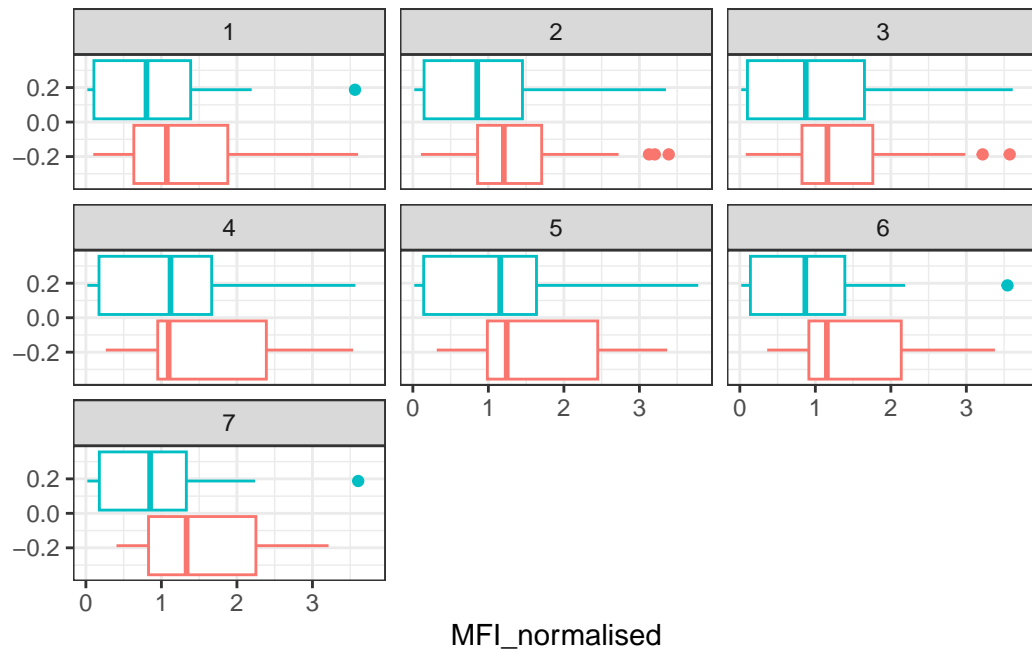
Q15

```
igg %>% filter(visit != 8) %>%
ggplot() +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  xlim(0,75) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```

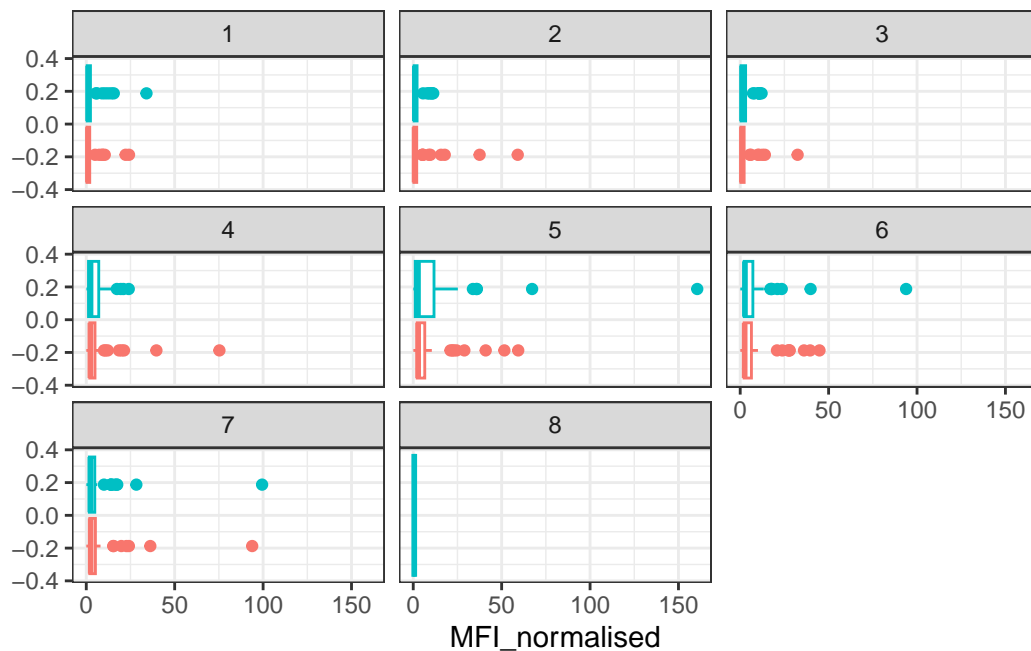
Warning: Removed 5 rows containing non-finite values (`stat_boxplot()`).



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



The boxplot above filters out two of the antigens for analysis, I picked OVA and FHA.

Q16

The antigen levels of FHA exceed the levels of OVA and continue to rise overtime. FHA peaks at 7, whereas OVA peaks at 5.

Q17 Both aP and wP have different antigen levels of OVA and FHA at different times. They are filtered out with different colors.