

Class 18 Pertussis mini Project

Nancy Leon-Rivera PID A12925462

Pertussis (a.k.a whooping cough) is a serious lung infection caused by the bacteria *B. Pertussis*.

The CDC tracks Pertussis case numbers in the US. The data is available at the following URL: <http://tinyurl.com/pertussiscdc>

We can “scrape” this data using the **datapasta** package.

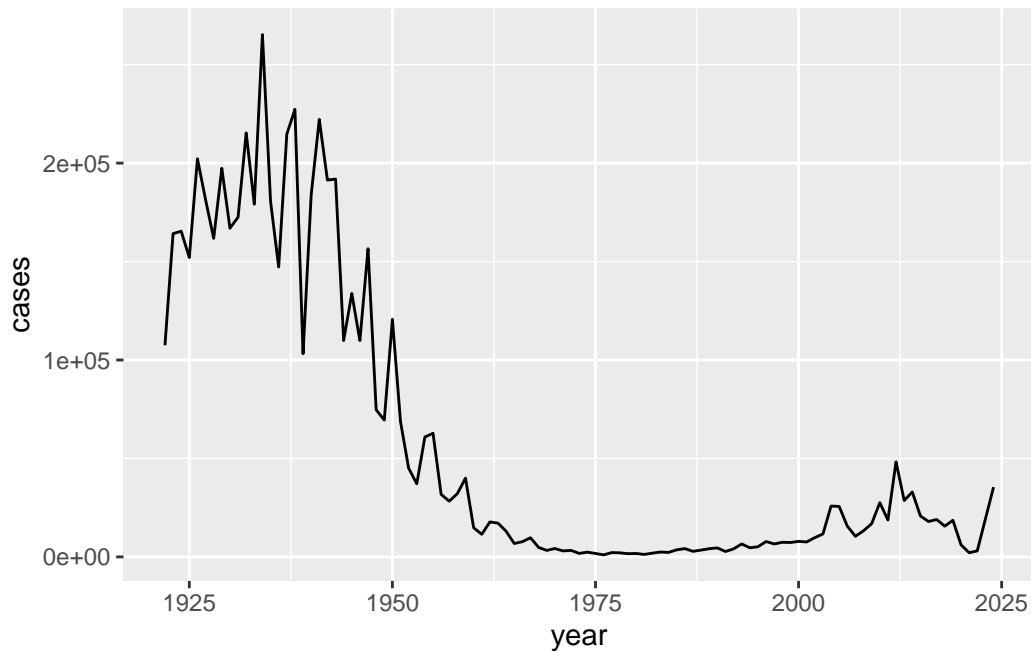
```
head(cdc)
```

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

Q1. Make a plot of pertussis cases per year using ggplot

```
library(ggplot2)

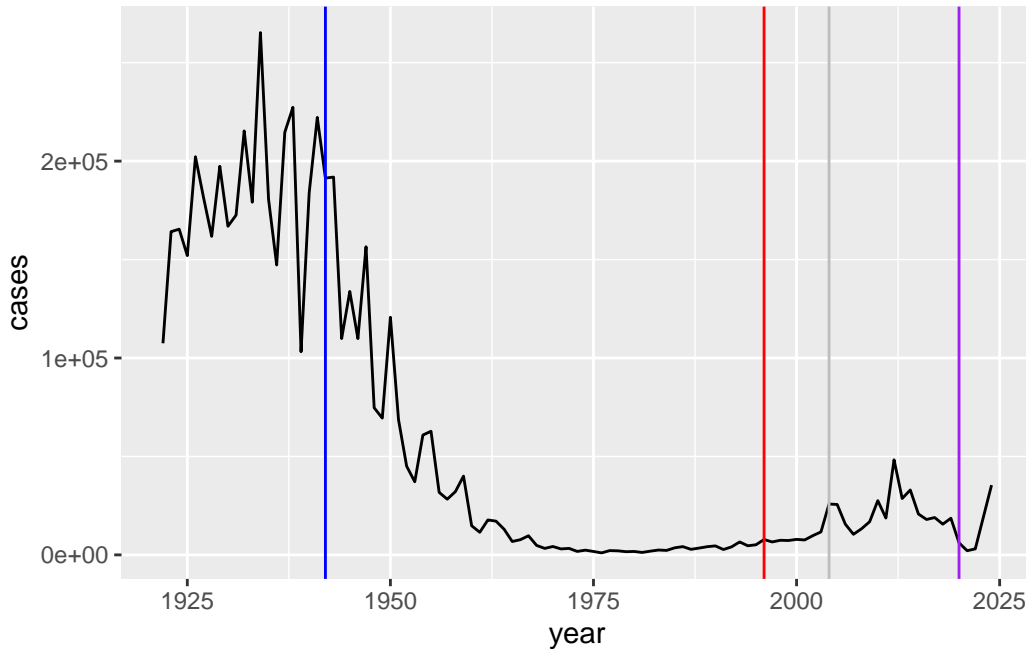
ggplot(cdc) +
  aes(x=year, y=cases) +
  geom_line()
```



Q2. Let's add key milestones of DTP (wP) vaccine roll out in 1942 and the switch to the new aP vaccine in 1996. We can use `geom_vline()` for this. Booster shots started in 2004

```
library(ggplot2)

ggplot(cdc) +
  aes(x=year, y=cases) +
  geom_line() +
  geom_vline(xintercept = 1942, col="blue")+
  geom_vline(xintercept = 1996, col = "red")+
  geom_vline(xintercept = 2020, col="purple") +
  geom_vline(xintercept = 2004, col="grey")
```



There are high case numbers pre 1946 (before the wP vaccine) then relatively rapid decrease in case numbers through the 1970s and on to 2004 when our first wodespread outbreak occured again.

Mounting evidence indicates that the aP vaccine induced immunity wanes faster than the older xP vaccine.

Enter the CMI-PB Project

CComputational Models of Immunity Pertussis Boost

One of the main goals of this project is to determine what is different in the immune response between wP and aP primed individuals

Using the booster vaccine as a proxy for infection.

All data is available here: <https://www.cmi-pb.org/> in JSON format.

```
library(jsonlite)
```

Warning: package 'jsonlite' was built under R version 4.4.3

```
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

Q. How many individuals “subjects” are there in this dataset?

```
nrow(subject)
```

```
[1] 172
```

Q. How many aP and wP subject are there?

```
table(subject$infacy_vac)
```

```
< table of extent 0 >
```

Q. Male/Female numbers

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

Female	Male
112	60

Q. breakdown of biological sex and race?

```
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

Q. Does this look to be representative of the US population at large?

No.

Let's read some more CMI-PB data

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

```
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

To use this data we need to *join* the various tables to find all the information I need to know about a particular measurement.

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

Now we can koin meta with ab_titer data.

```
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	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

Q. How many different antibody isotypes are we measuring?

```
table(ab_data$isotype)
```

IgE	IgG	IgG1	IgG2	IgG3	IgG4
6698	7265	11993	12000	12000	12000

Q. How many different antigens?


```
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				

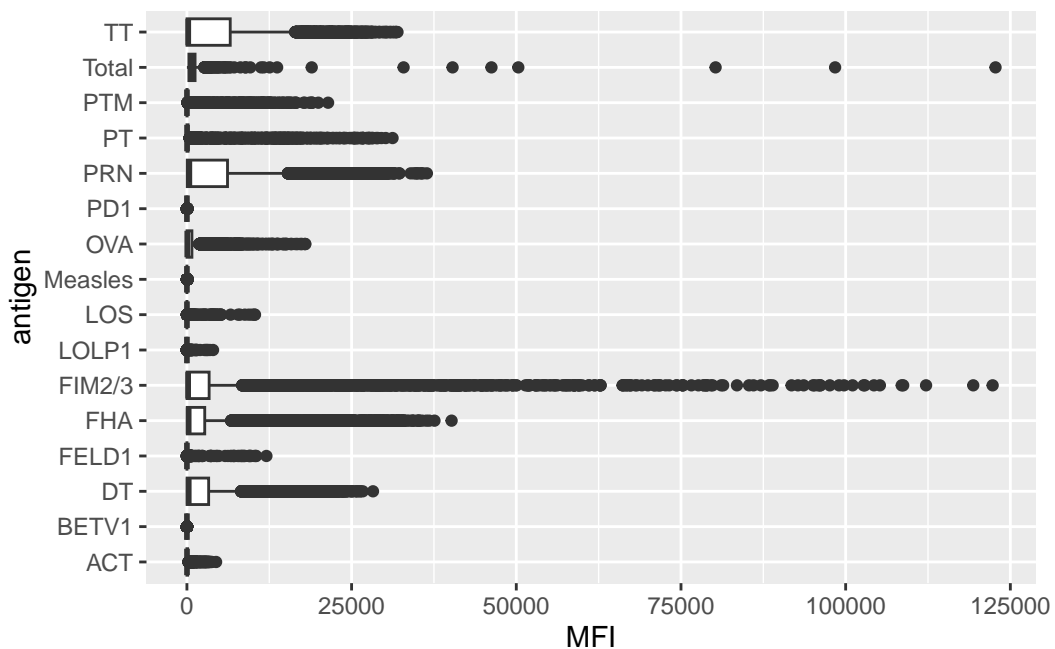
Q. Let's look at a box plot of antigen levels over the whole dataset?

```
dim(ab_data)
```

```
[1] 61956    20
```

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

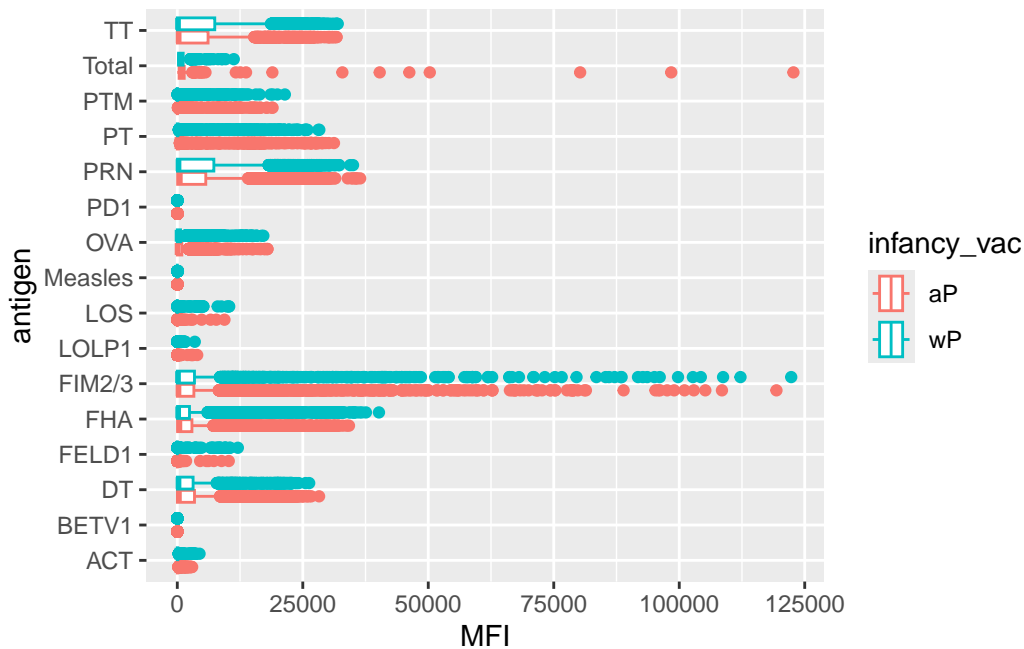
Warning: Removed 1 row containing non-finite outside the scale range (`stat_boxplot()`).



Q. Break this plot down by aP or wP. Can have different color for each

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

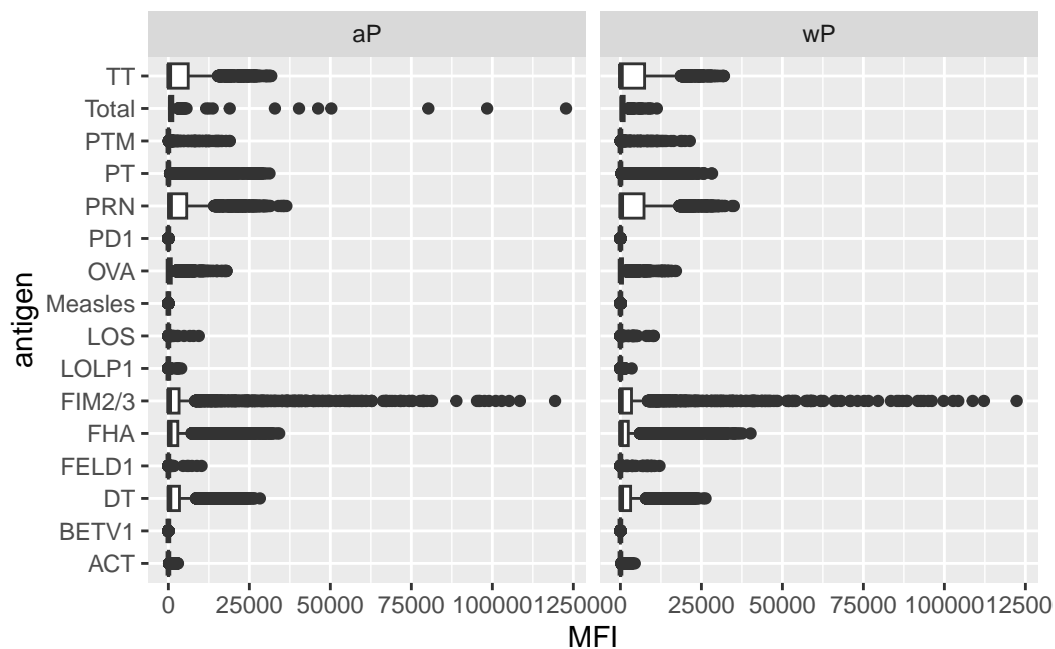
Warning: Removed 1 row containing non-finite outside the scale range
(`stat_boxplot()`).



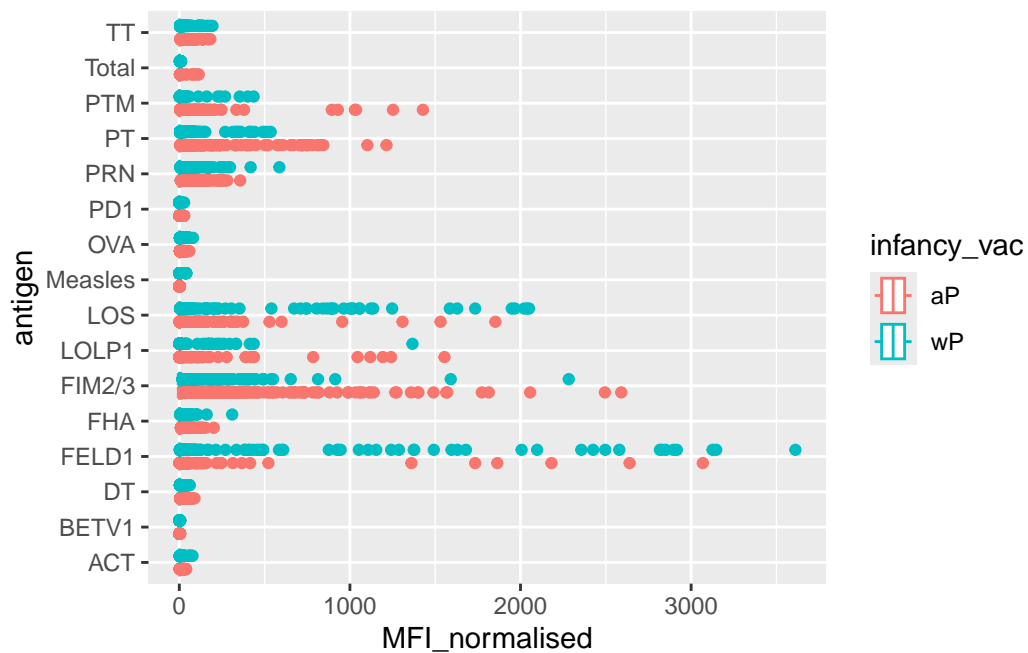
We can facet the plot by infancy_vac.

```
ggplot(ab_data)+  
  aes(MFI, antigen)+  
  geom_boxplot() +  
  facet_wrap(~infancy_vac)
```

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



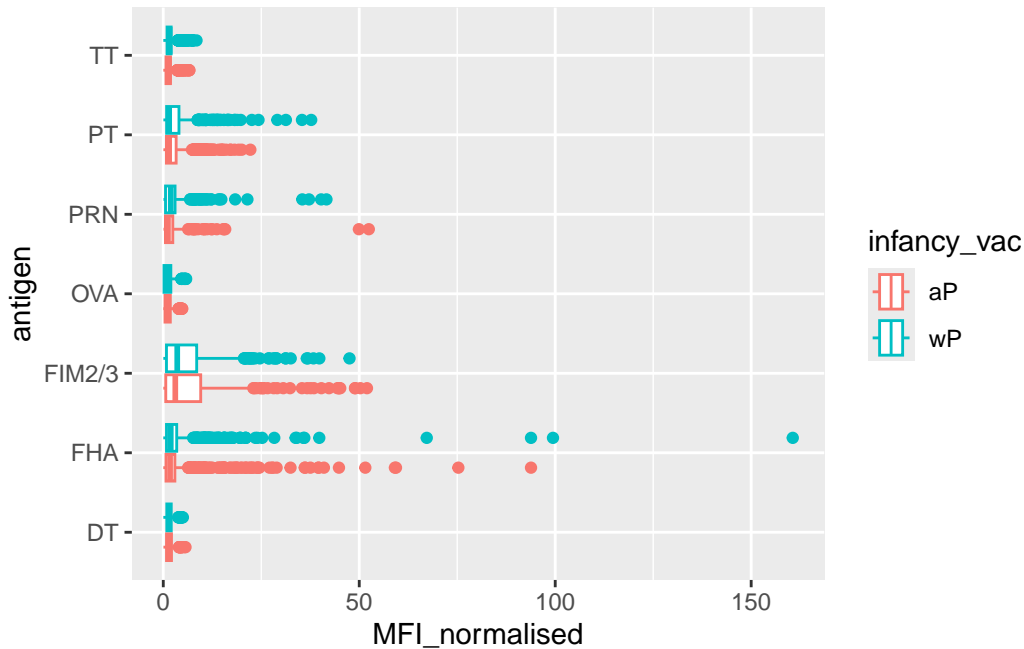
```
ggplot(ab_data)+
  aes(MFI_normalised, antigen, col=infancy_vac)+
  geom_boxplot()
```



Let's focus on just IgG

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

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

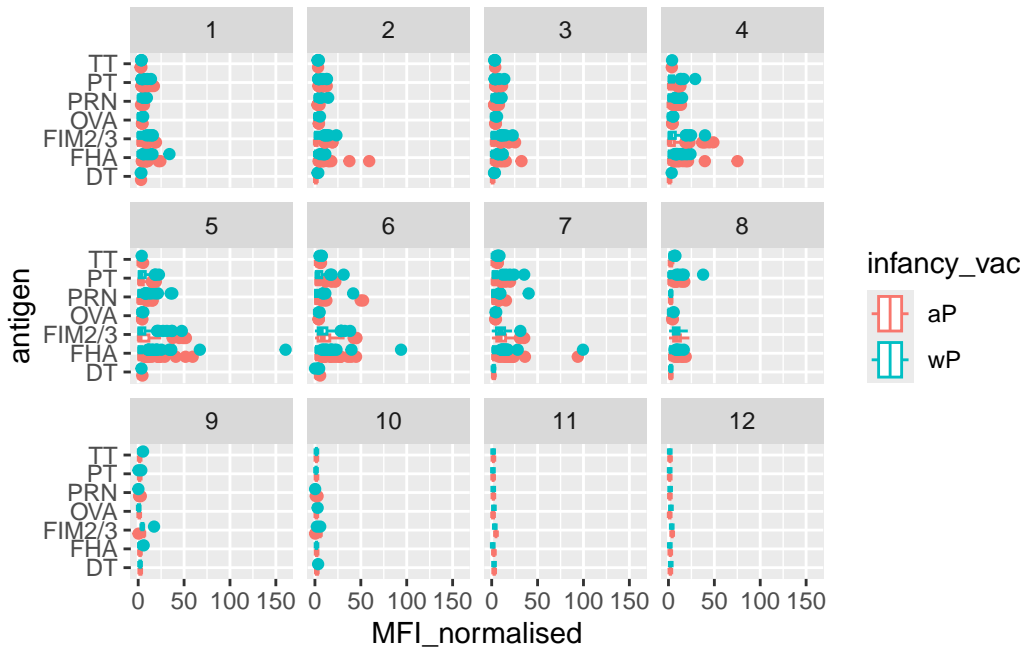


```
head(igg)
```

	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						

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



Let's focus on PT(petussis Toxin) and igg levels over time

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

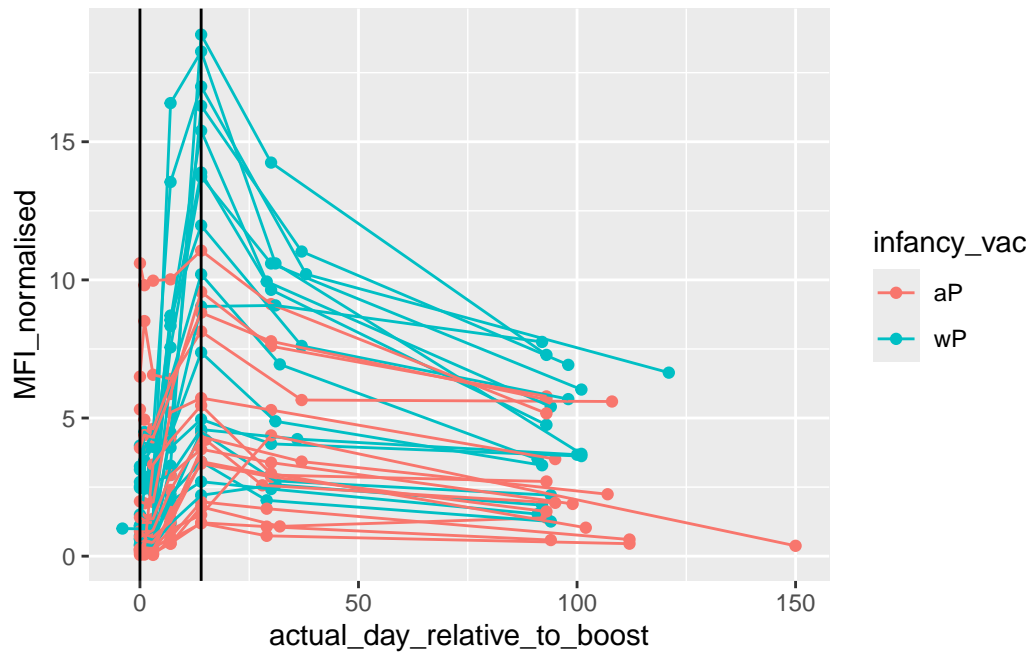
2020_dataset	2021_dataset	2022_dataset	2023_dataset
31520	8085	7301	15050

Filter to focus on one antigen, PT, and IgG levels

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

A plot of actual days relative to boost vs MFI _normalised

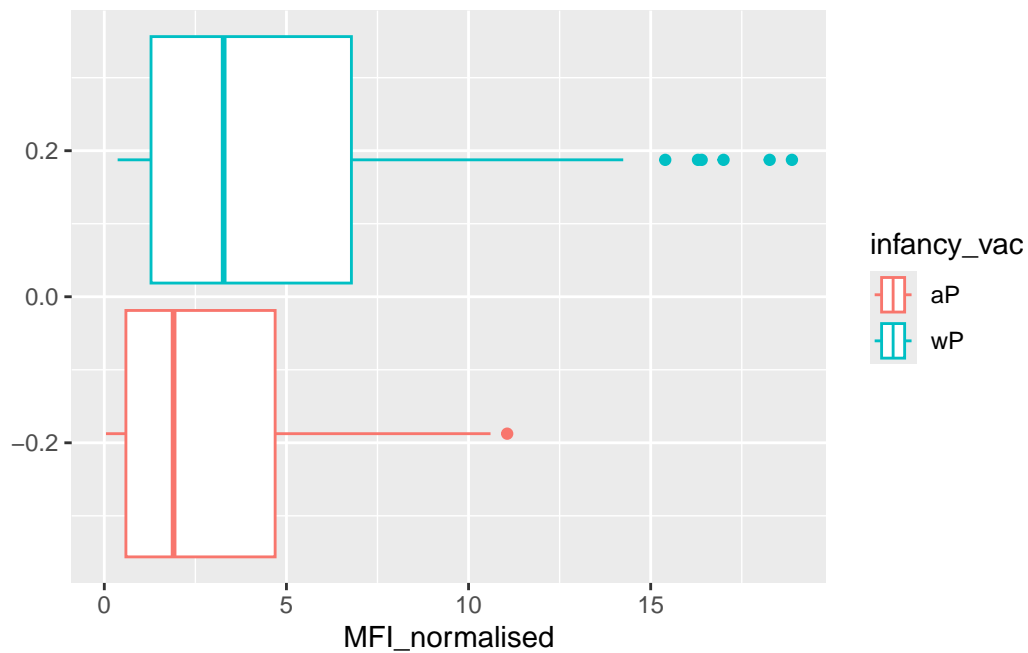
```
ggplot(pt_igg)+
  aes(actual_day_relative_to_boost, MFI_normalised, col=infancy_vac, group = subject_id)+
  geom_point() +
  geom_line() +
  geom_vline(xintercept = 14) +
  geom_vline(xintercept = 0)
```



```
labs(title = "IgG PT Levels",
     xlab = "Days relative to Boost") +
geom_smooth(method = "loess", se = T)
```

NULL

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



```
library(lubridate)
```

Attaching package: 'lubridate'

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

date, intersect, setdiff, union

```
#subject$year_of_birth
today()
```

```
[1] "2025-03-07"
```

```
today() - mdy ("04-02-1997")
```

Time difference of 10201 days


```
today() - dmy()
```

difftime of length 0

```
time_length(today() - mdy ("04-02-1997"), "years")
```

```
[1] 27.92882
```

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