

Mini-Project Investigating Pertussis Resurgence

AUTHOR

Monika Ramos

Class 18 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.

#Is pertussis on the rise?

```
library(datapasta)

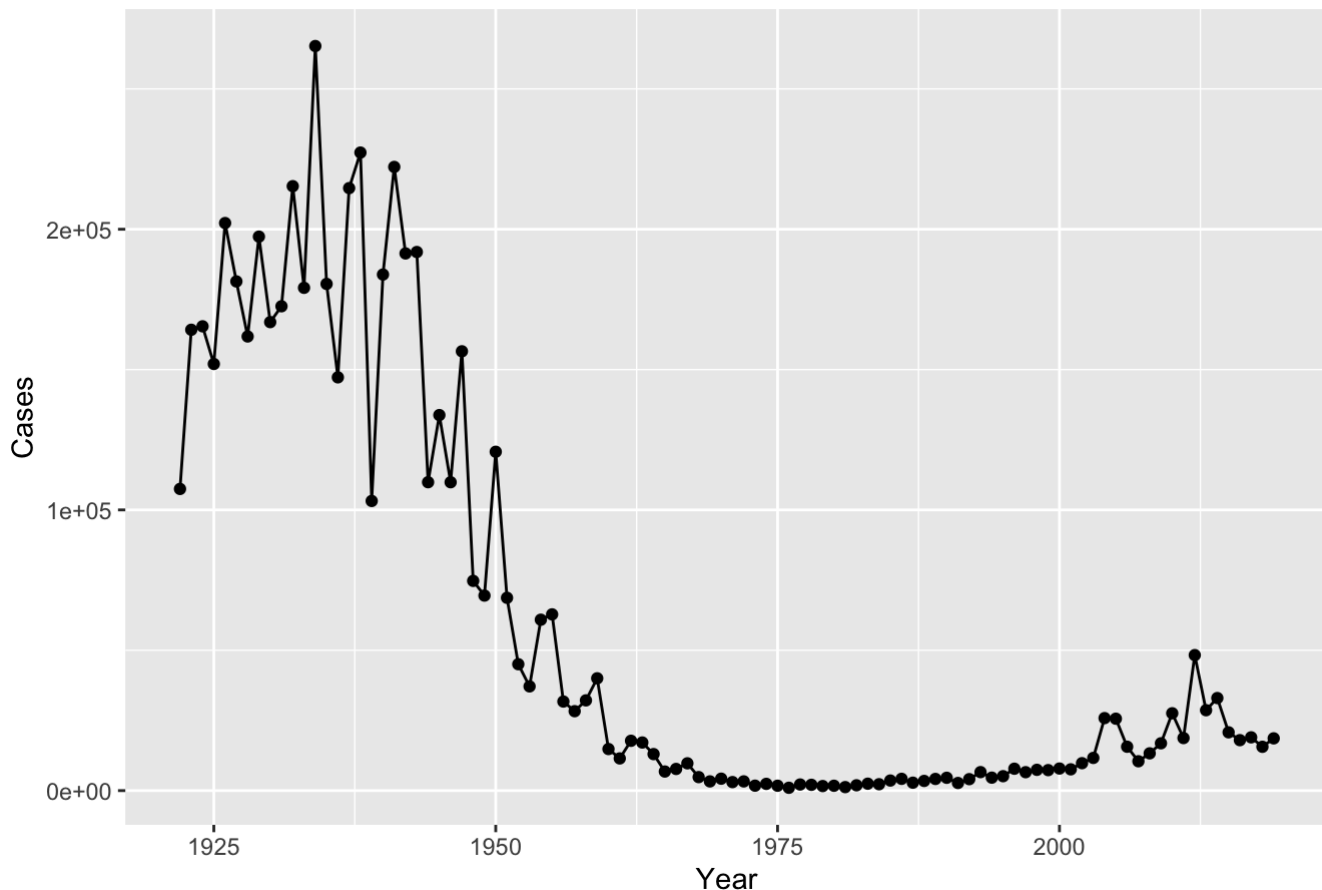
#use datapasta to paste copied CDC data table and paste it as dataframe.
cdc <- data.frame(
  Year = c(1922L,
            1923L, 1924L, 1925L, 1926L, 1927L, 1928L,
            1929L, 1930L, 1931L, 1932L, 1933L, 1934L, 1935L,
            1936L, 1937L, 1938L, 1939L, 1940L, 1941L,
            1942L, 1943L, 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),
  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)
)
```

Plot the data

Plot the data

```
library(ggplot2)
ggplot(data = cdc) + aes(x=Year, y=Cases) +
  geom_point() +
  geom_line() +
  labs(title = "Pertussis cases per year in the USA")
```

Pertussis cases per year in the USA



```
baseplot <- ggplot(data = cdc) + aes(x=Year, y=Cases) +
  geom_point() +
  geom_line() +
  labs(title = "Pertussis cases per year in the USA")
```

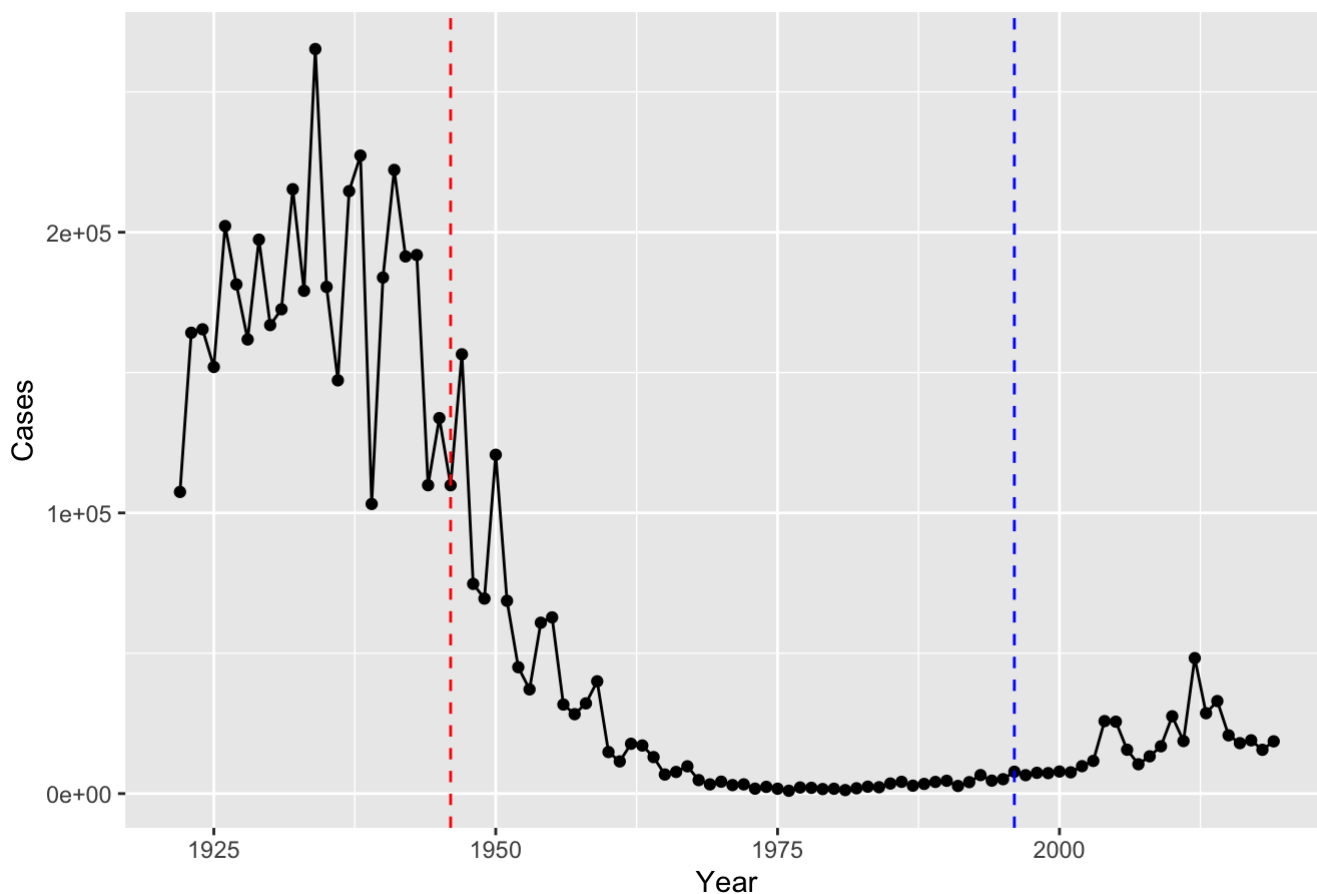
Two types of pertussis vaccines have been developed: whole-cell pertussis (wP) and acellular pertussis (aP). The first vaccines were composed of 'whole cell' (wP) inactivated bacteria. The latter aP vaccines use purified antigens of the bacteria (the most important pertussis components for our immune system). These aP vaccines were developed to have less side effects than the older wP vaccines and are now the only form administered in the United States.

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?

```
baseplot +
  geom_vline(xintercept = 1946, color = "red", linetype = 2) +
```

```
geom_vline(xintercept = 1946, color = "red", linetype = 2)
geom_vline(xintercept = 1996, color = "blue", linetype = 2)
```

Pertussis cases per year in the USA



There is a increase or resurgence after year 2000.

Q3. Describe what happened after the introduction of the aP vaccine? Do you have a possible explanation for the observed trend? A few years after the introduction of the aP vaccine, there is an evident surge in pertussis cases. I think a few possible explanations for this could be that the aP vaccine may not contain the necessary antigens for effective vaccination against *Bordetella pertussis* in comparison to the whole cell vaccine. Additionally, perhaps the pathogen mutated enough to overcome the immunization efficacy and be different enough that the vaccination is no longer efficacious.

- 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.

The new and ongoing CMI-PB project aims to provide the scientific community with this very information. In particular, CMI-PB tracks and makes freely available long-term humoral and cellular immune response data for a large number of individuals who received either DTwP or DTaP combination vaccines in infancy followed by Tdap booster vaccinations. This includes complete API access to longitudinal RNA-Seq, AB Titer, Olink, and live cell assay results directly from their website: <https://www.cmi-pb.org/>

```
# Allows us to read, write and process JSON data
library(jsonlite)
subject <- read_json("https://www.cmi-pb.org/api/subject", simplifyVector = TRUE)
head(subject, 3)
```

```
  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
  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
```

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

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

```
aP wP
47 49
```

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

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

```
Female  Male
   66    30
```

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	18	9
Black or African American	2	0
More Than One Race	8	2
Native Hawaiian or Other Pacific Islander	1	1
Unknown or Not Reported	10	4
White	27	13

Side-Note: Working with dates Two of the columns of subject contain dates in the Year-Month-Day format. Recall from our last mini-project that dates and times can be annoying to work with at the best

of times. However, in R we have the excellent lubridate package, which can make life a lot easier. Here is a quick example to get you started:

```
library(lubridate)
```

Loading required package: timechange

Attaching package: 'lubridate'

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

date, intersect, setdiff, union

```
today()
```

```
[1] "2022-12-04"
```

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

Time difference of 8373 days

```
time_length( today() - ymd("2000-01-01"), "years")
```

```
[1] 22.92402
```

Note that here we are using the ymd() function to tell lubridate the format of our particular date and then the time_length() function to convert days to 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?

```
subject$age <- time_length(today() - ymd(subject$year_of_birth), "years")
#what is the average age of all subjects?
mean(subject$age)
```

```
[1] 30.77701
```

```
summary(subject$age)
```

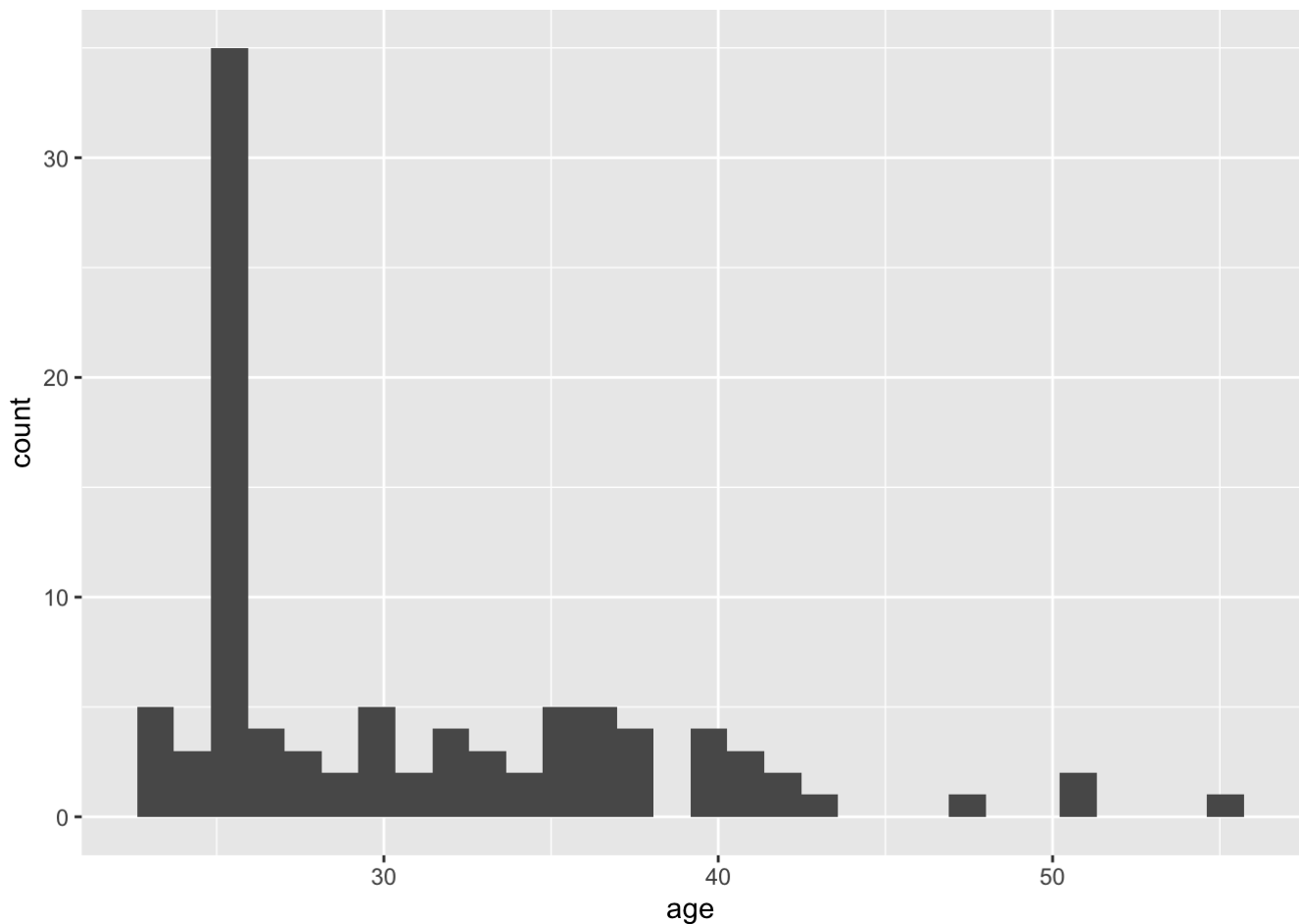
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
22.92	25.92	27.92	30.78	34.92	54.92

Histogram

```
ggplot(subject) +
  aes(age) +
```

```
aes(age, )  
geom_histogram()
```

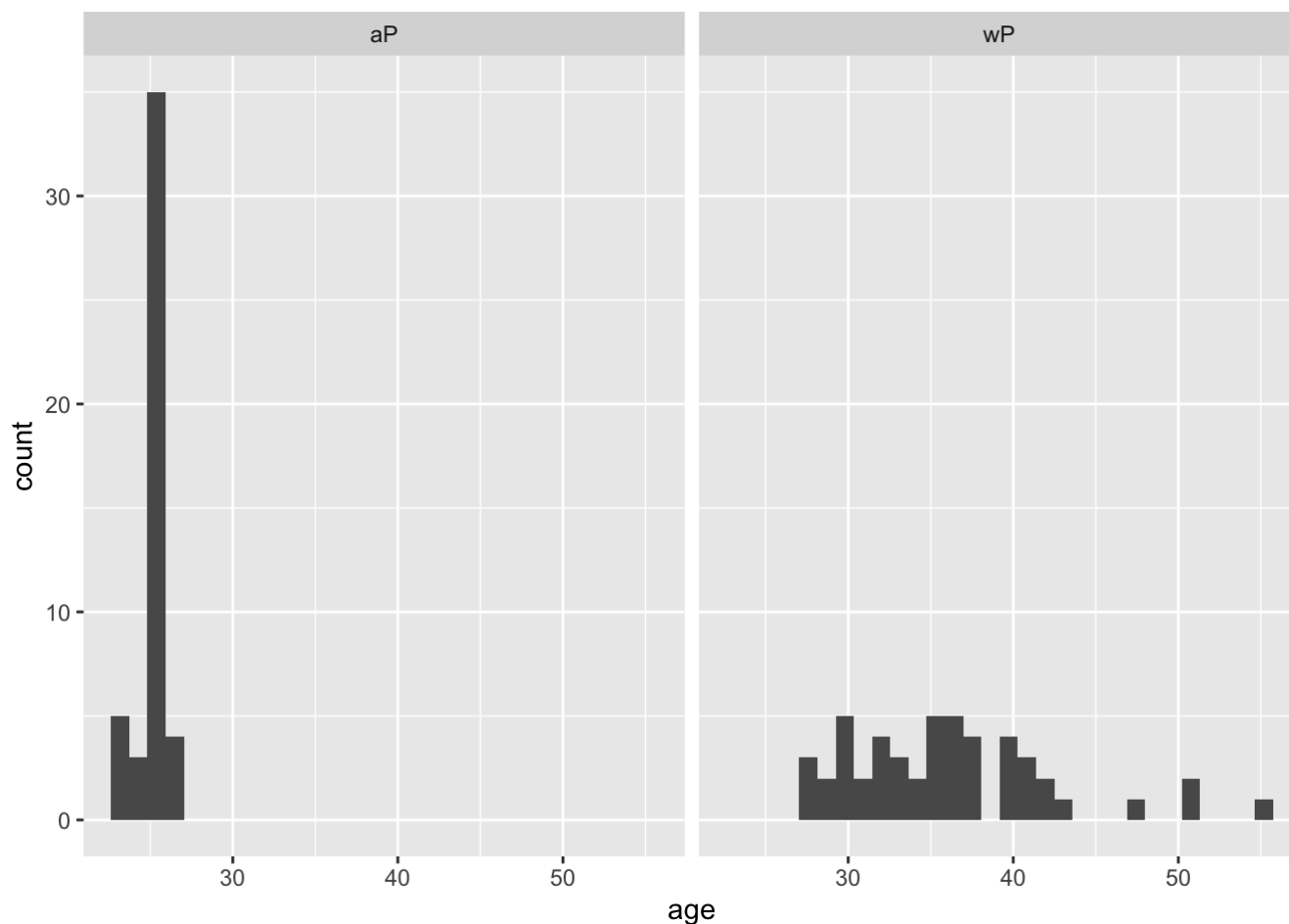
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.



There is an evident peak in younger age groups.

```
#add a faceted wrap layer by wP and aP  
ggplot(subject) +  
  aes(age) +  
  geom_histogram() +  
  facet_wrap(vars(infancy_vac))
```

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



Yes they are different.

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

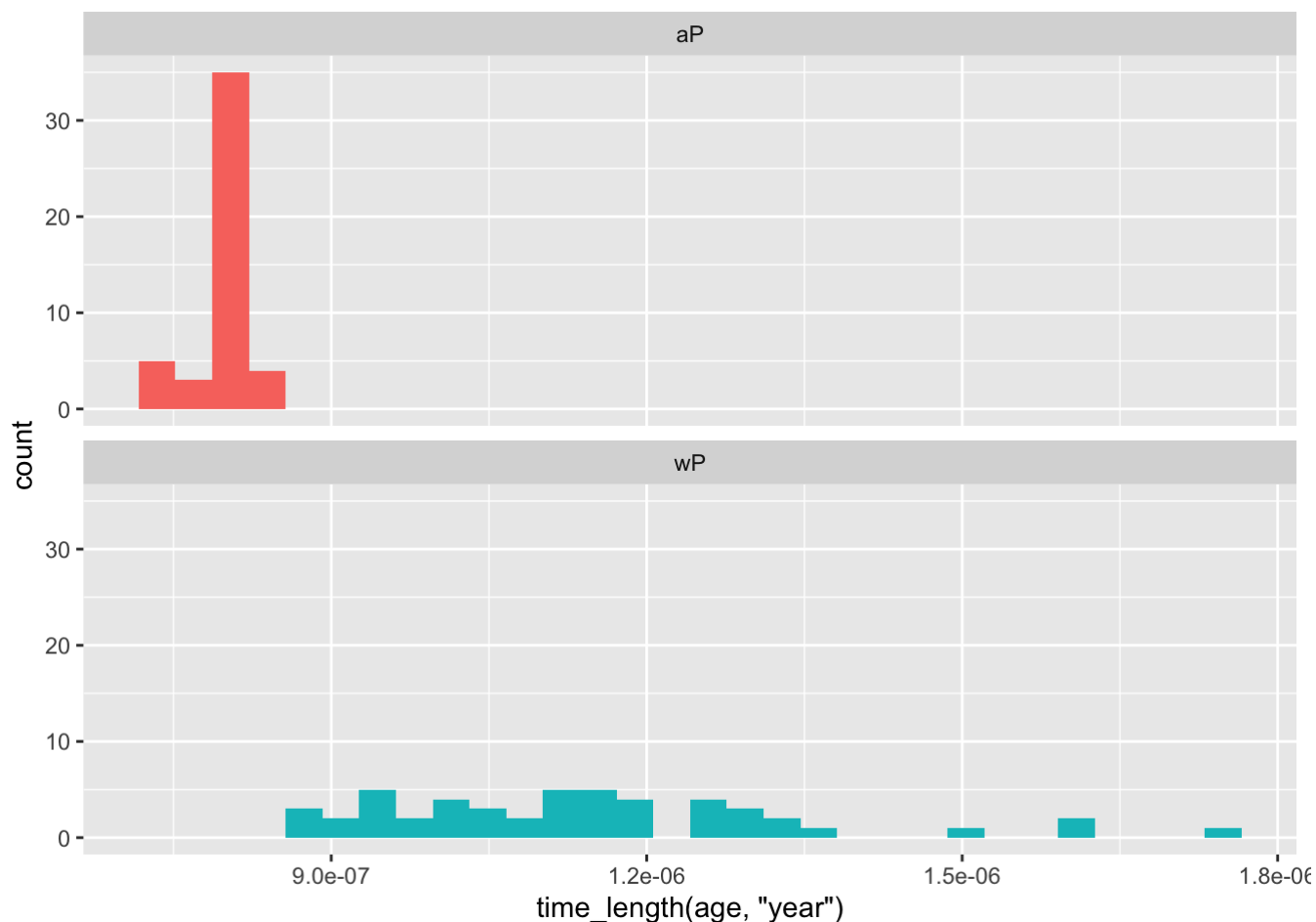
```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

```
[1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481
```

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)
```

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



Yes they appear very different.

Joining multiple tables Read the specimen and ab_titer tables into R and store the data as specimen and titer named data frames.

```
# Complete the API URLs...
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	736	
3	3	1	1	
4	4	1	3	
5	5	1	7	
6	6	1	11	

	planned_day_relative_to_boost	specimen_type	visit
1	0	Blood	1
2	736	Blood	10
3	1	Blood	2
4	3	Blood	3
5	7	Blood	4
6	14	Blood	5

We want to join the 'subject' and the 'specimen' tables to have all the metadata we need for later analysis. we can use the dplyr *_join() funtions for this task.

```
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(specimen,subject)
```

Joining, by = "subject_id"

```
dim(meta)
```

```
[1] 729 14
```

Now read antibody titre experimental data

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

I need to linke a.k.a join the ab titre data with our meta data

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

Joining, by = "specimen_id"

```
dim(abdata)
```

```
[1] 32675    21
```

This is getting really large.

4. Examine IgG1 Ab titer levels

```
ig1 <- abdata %>% filter(isotype == "IgG1", visit!=8)
head(ig1)
```

	specimen_id	isotype	is_antigen_specific	antigen	MFI	MFI_normalised
1	1	IgG1	TRUE	ACT	274.355068	0.6928058
2	1	IgG1	TRUE	LOS	10.974026	2.1645083
3	1	IgG1	TRUE	FELD1	1.448796	0.8080941
4	1	IgG1	TRUE	BETV1	0.100000	1.0000000
5	1	IgG1	TRUE	LOLP1	0.100000	1.0000000
6	1	IgG1	TRUE	Measles	36.277417	1.6638332

	unit	lower_limit_of_detection	subject_id	actual_day_relative_to_boost
1	IU/ML	3.848750	1	-3
2	IU/ML	4.357917	1	-3
3	IU/ML	2.699944	1	-3
4	IU/ML	1.734784	1	-3
5	IU/ML	2.550606	1	-3
6	IU/ML	4.438966	1	-3

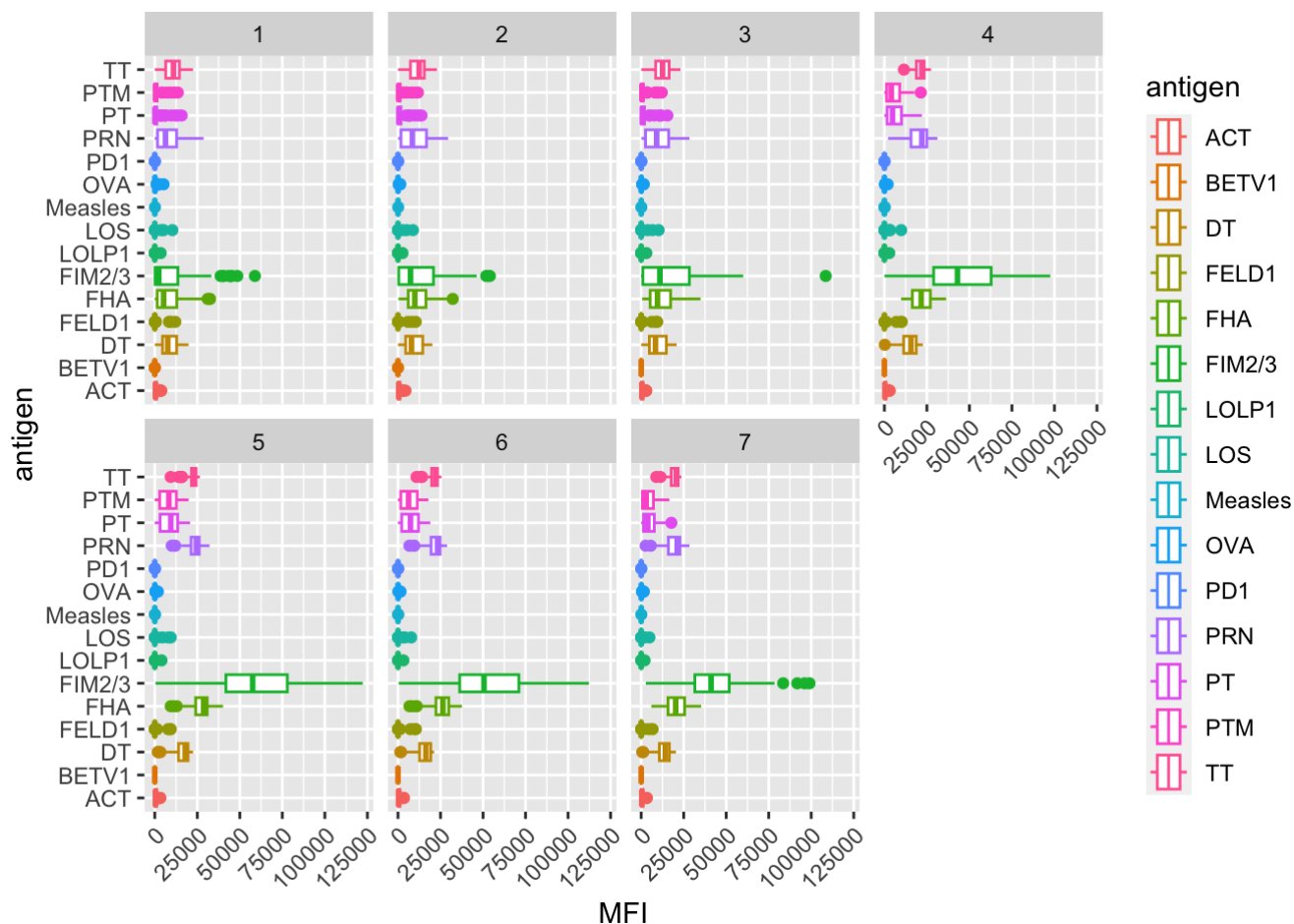
	planned_day_relative_to_boost	specimen_type	visit	infancy_vac	biological_sex
1	0	Blood	1	wP	Female
2	0	Blood	1	wP	Female
3	0	Blood	1	wP	Female
4	0	Blood	1	wP	Female
5	0	Blood	1	wP	Female
6	0	Blood	1	wP	Female

	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
1	36.92266
2	36.92266
3	36.92266
4	36.92266
5	36.92266
6	36.92266

Q13. Complete the following code to make a summary boxplot of Ab titer levels for all antigens:

```
ggplot(ig1) +
  aes(MFI, antigen, color= antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow=2) +
  theme(axis.text.x = element_text(angle = 45, hjust=1))
```



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

```
IgE  IgG  IgG1  IgG2  IgG3  IgG4
6698 1413 6141 6141 6141 6141
```

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

```
1    2    3    4    5    6    7    8
5795 4640 4640 4640 4640 4320 3920 80
```

4. Examine IgG1 Ab titer levels

```
ig1 <- abdata %>% filter(isotype == "IgG1", visit!=8)
head(ig1)
```

	specimen_id	isotype	is_antigen_specific	antigen	MFI	MFI_normalised
1	1	IgG1	TRUE	ACT	274.355068	0.6928058
2	1	IgG1	TRUE	LOS	10.974026	2.1645083
3	1	IgG1	TRUE	FELD1	1.448796	0.8080941
4	1	IgG1	TRUE	BETV1	0.100000	1.0000000
5	1	IgG1	TRUE	LOLP1	0.100000	1.0000000
6	1	IgG1	TRUE	Measles	36.277417	1.6638332

	unit	lower_limit_of_detection	subject_id	actual_day_relative_to_boost
1	IU/ML	3.848750	1	-3
2	IU/ML	4.357917	1	-3
3	IU/ML	2.699944	1	-3
4	IU/ML	1.734784	1	-3
5	IU/ML	2.550606	1	-3
6	IU/ML	4.438966	1	-3

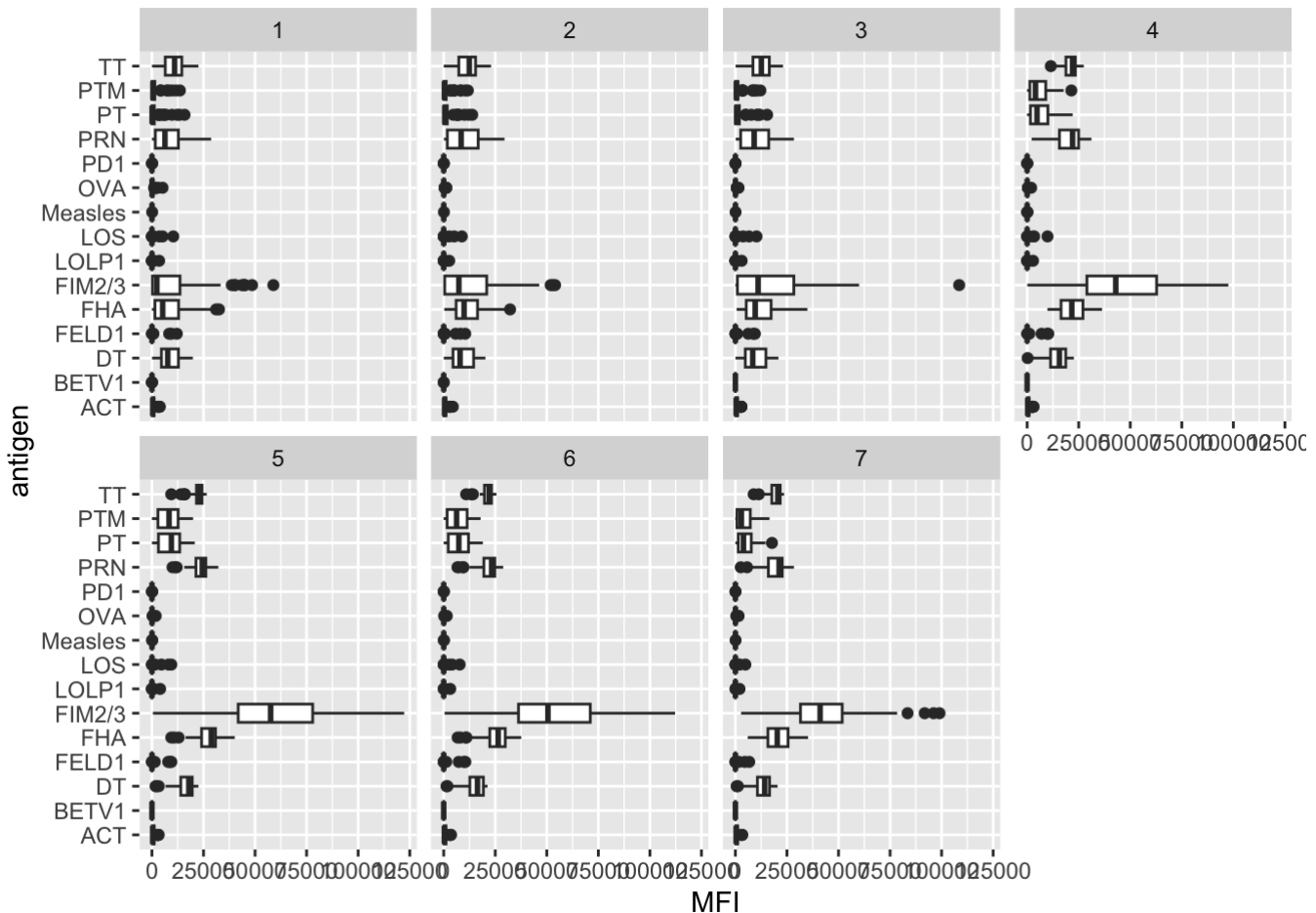
	planned_day_relative_to_boost	specimen_type	visit	infancy_vac	biological_sex
1	0	Blood	1	wP	Female
2	0	Blood	1	wP	Female
3	0	Blood	1	wP	Female
4	0	Blood	1	wP	Female
5	0	Blood	1	wP	Female
6	0	Blood	1	wP	Female

	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
1	36.92266
2	36.92266
3	36.92266
4	36.92266
5	36.92266
6	36.92266

Q13. Complete the following code to make a summary boxplot of Ab titer levels for all antigens:

```
ggplot(ig1) +
  aes(MFI, antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow=2)
```

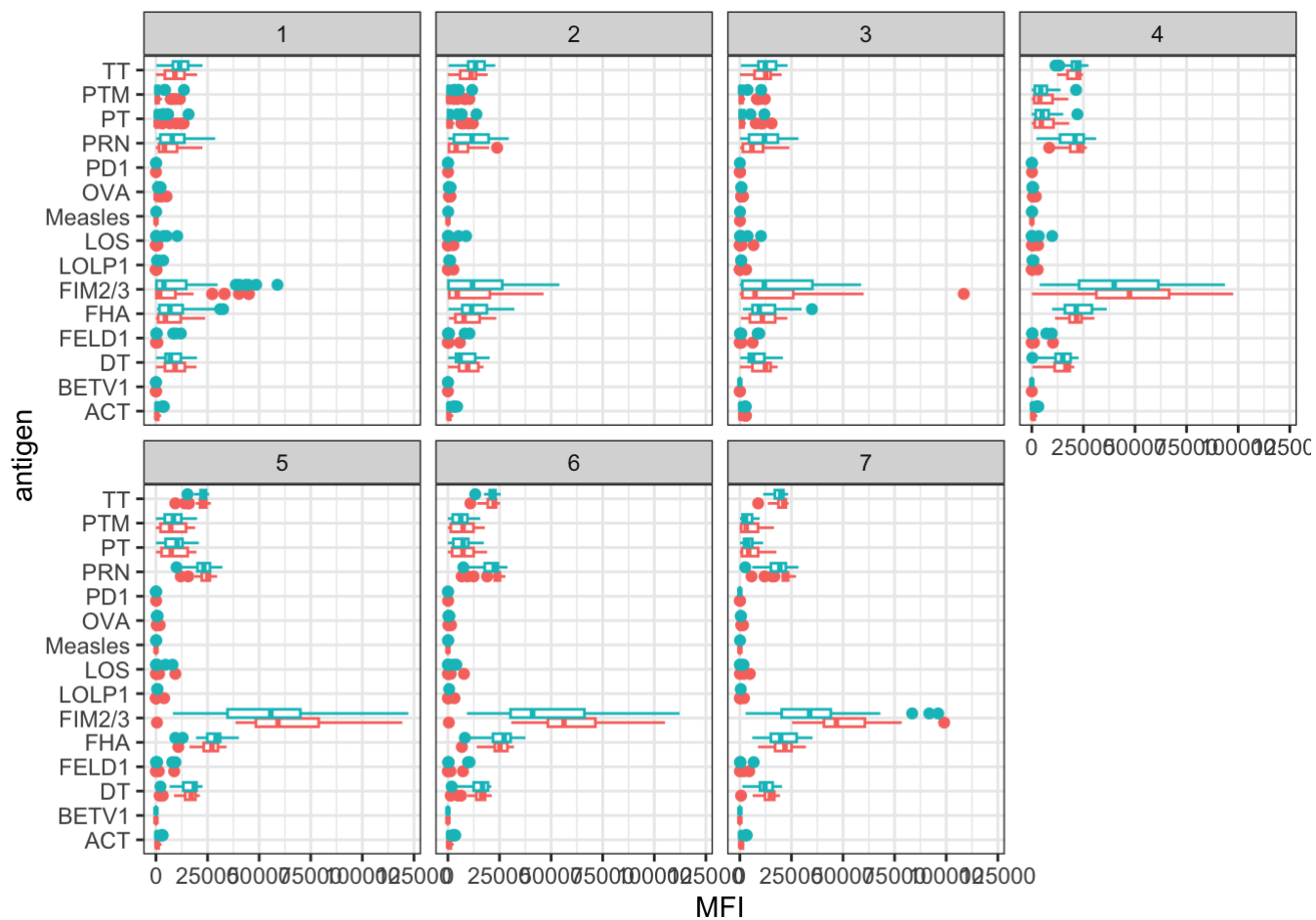


Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

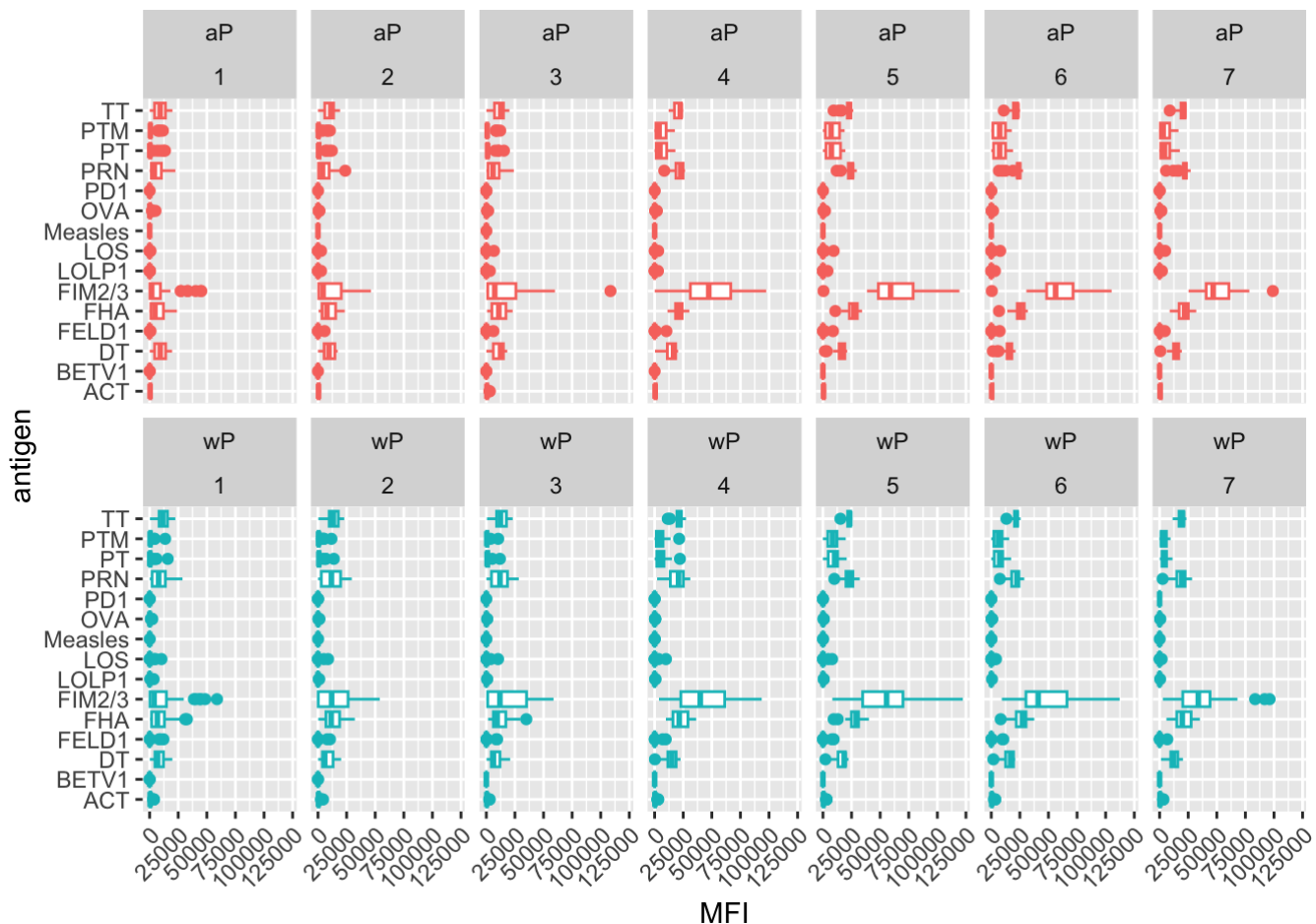
It looks like FIM2/3 has the most IgG1 antibody titre and this could be because it is an immunodominant antigen. Perhaps it is more antibody-accessible than the others as well. Lastly, it could be extracellular in contrast to intracellular antigens which require MHC-I or MHC-II antigen processing and presentation.

We can attempt to examine differences between wP and aP here by setting color and/or facet values of the plot to include infancy_vac status (see below). However these plots tend to be rather busy and thus hard to interpret easily.

```
ggplot(ig1) +
  aes(MFI, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  theme_bw()
```

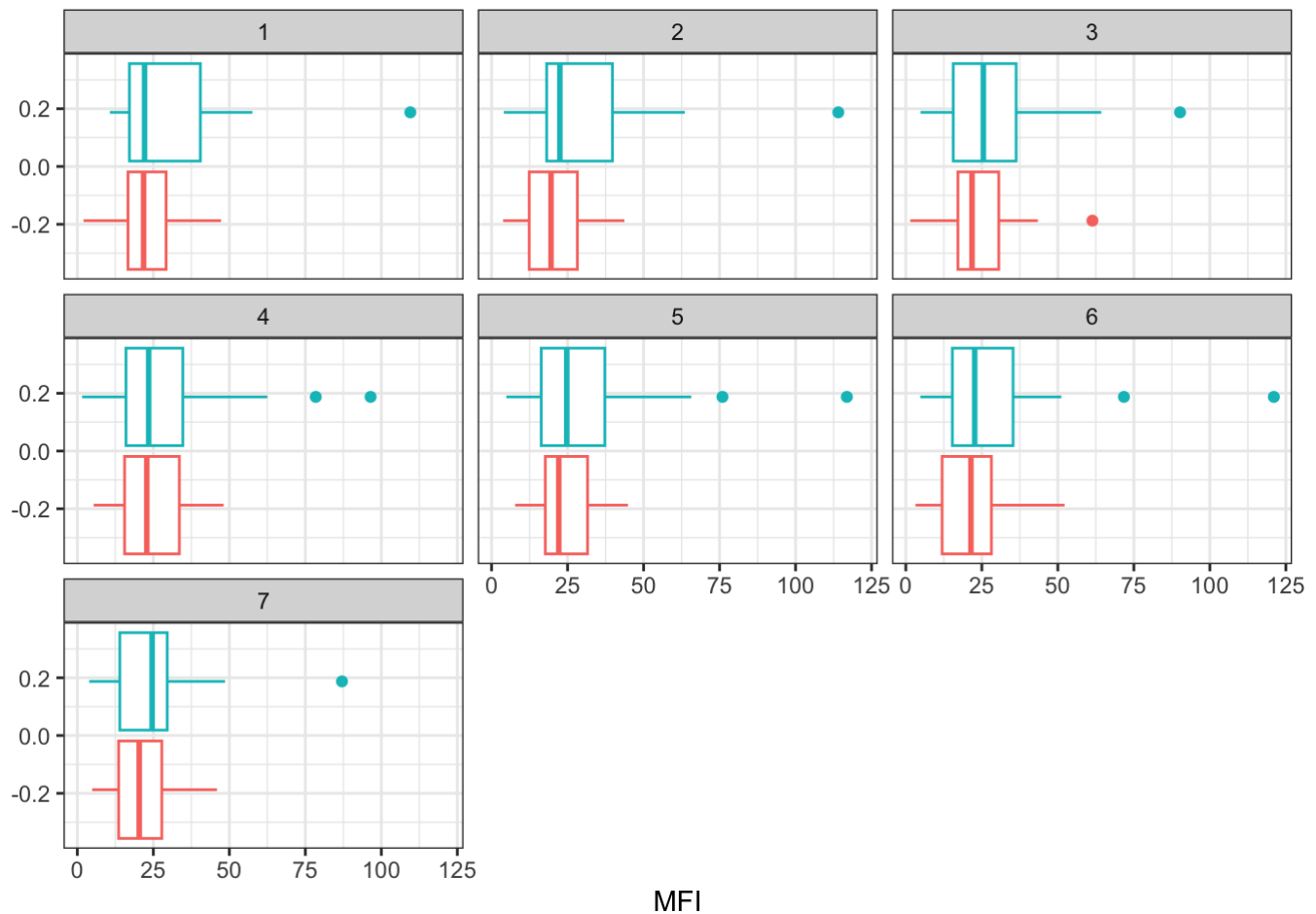


```
ggplot(ig1) +
  aes(MFI, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(infancy_vac, visit), nrow=2) +
  theme(axis.text.x = element_text(angle = 45, hjust=1))
```



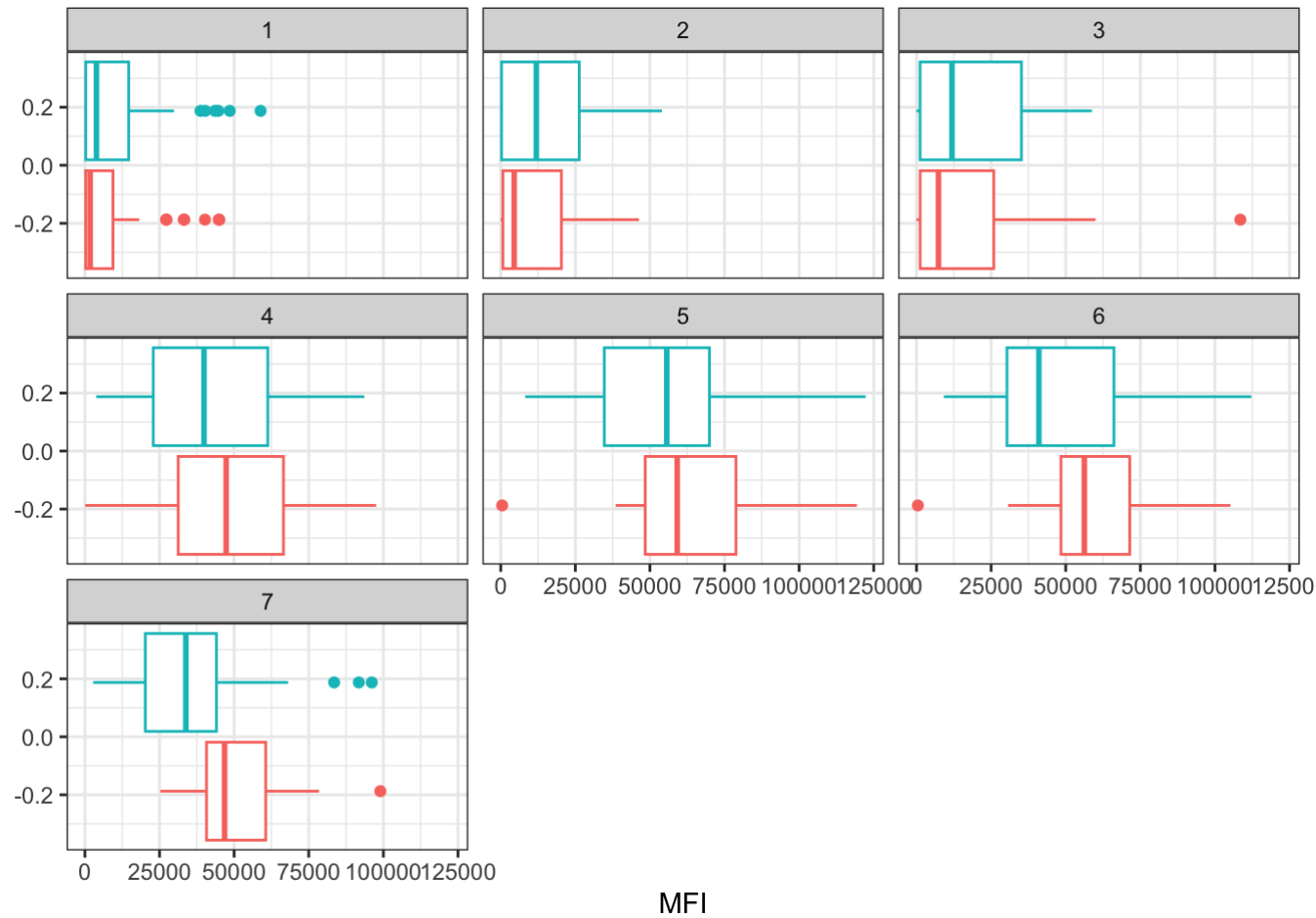
Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can choose any you like. Below I picked a "control" antigen ("Measles", that is not in our vaccines) and a clear antigen of interest ("FIM2/3", extra-cellular fimbriae proteins from *B. pertussis* that participate in substrate attachment).

```
filter(ig1, antigen=="Measles") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = F) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



and the same for antigen=="FIM2/3"

```
filter(ig1, antigen=="FIM2/3") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = F) +
  facet_wrap(vars(visit)) +
  theme_bw()
```

MFI