

# class 15

Gaeun Jun (A16814573)

## 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.

```
library(datapasta)
library(ggplot2)
```

```
cdc <- data.frame(
  Year = c(1922, 1923, 1924, 1925, 1926, 1927, 1928,
    1929, 1930, 1931, 1932, 1933, 1934, 1935,
    1936, 1937, 1938, 1939, 1940, 1941,
    1942, 1943, 1944, 1945, 1946, 1947, 1948,
    1949, 1950, 1951, 1952, 1953, 1954,
    1955, 1956, 1957, 1958, 1959, 1960,
    1961, 1962, 1963, 1964, 1965, 1966, 1967,
    1968, 1969, 1970, 1971, 1972, 1973,
    1974, 1975, 1976, 1977, 1978, 1979, 1980,
    1981, 1982, 1983, 1984, 1985, 1986,
    1987, 1988, 1989, 1990, 1991, 1992, 1993,
    1994, 1995, 1996, 1997, 1998, 1999,
    2000, 2001, 2002, 2003, 2004, 2005,
    2006, 2007, 2008, 2009, 2010, 2011, 2012,
    2013, 2014, 2015, 2016, 2017, 2018,
    2019, 2020, 2021, 2022),
  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,
2116,3044)
)

colnames(cdc) <- c("Year", "Cases")

```

```

ggplot(cdc) +
  aes(x = Year, y = Cases) +
  geom_point(color = "black") +
  geom_line(color = "black") +
  labs(
    title = "Pertussis Cases Over Time",
    x = "Year",
    y = "Number of Cases"
  )

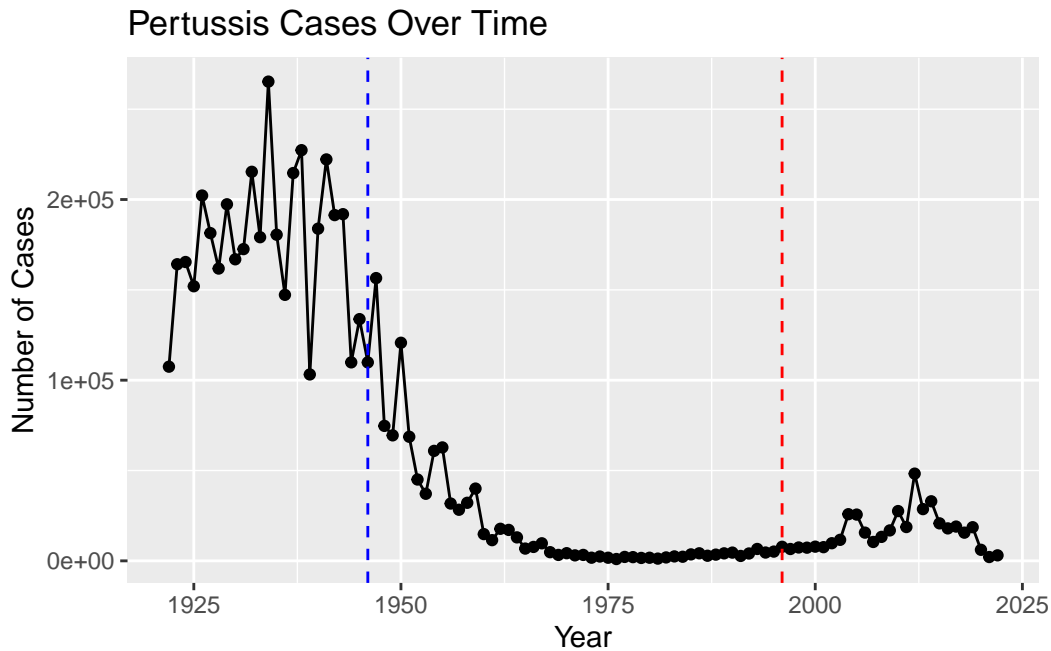
```



## 2. 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(color = "black") +
  geom_line(color = "black") +
  geom_vline(xintercept = 1946, linetype = "dashed", color = "blue") +
  geom_vline(xintercept = 1996, linetype = "dashed", color = "red") +
  labs(
    title = "Pertussis Cases Over Time",
    x = "Year",
    y = "Number of Cases"
  )
```



I noticed that the aP vaccine got introduced after the Pertussis cases have been declining (even with the wP vaccine).

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

After the introduction of the aP vaccine, the the number of cases stayed pretty consistently low. This trend could possibly be explained by the fact that this new vaccine has antigens that target the right bacteria. However, after a little bit, the cases start to rise again and this could be due to a growing immunization and resistance to the vaccine administration.

### 3. Exploring CMI-PB data

The CMI-PB API returns 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
87 85
```

There are 60 aP and 85 wP infancy vaccinated subjects in the dataset.

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

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

```
Female  Male
112     60
```

There are 112 females and 60 males in the dataset.

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	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

### Side-Note: Working with dates

```
library(lubridate)
```

Attaching package: 'lubridate'

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

date, intersect, setdiff, union

```
today()
```

```
[1] "2024-11-27"
```

```
# How many days have passed since new year 2000
today() - ymd("2000-01-01")
```

Time difference of 9097 days

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

```
[1] 24.90623
```

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?

```
# Use today's date to calculate age in days
subject$age <- today() - ymd(subject$year_of_birth)
subject$age_years <- time_length(subject$age, "years")
```

```
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

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

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

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
22	26	27	27	28	34

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

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
22	32	34	36	39	57

```
t.test(ap$age_years, wp$age_years)
```

Welch Two Sample t-test

```
data: ap$age_years and wp$age_years
t = -12.918, df = 104.03, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.094058 -7.407351
sample estimates:
mean of x mean of y
 26.80158  35.55228
```

Average age of wP individuals: 36 years Average age of aP individuals: 27 years Since the p-value is less than 0.05, the average ages of the wP and aP individuals are significantly different.

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

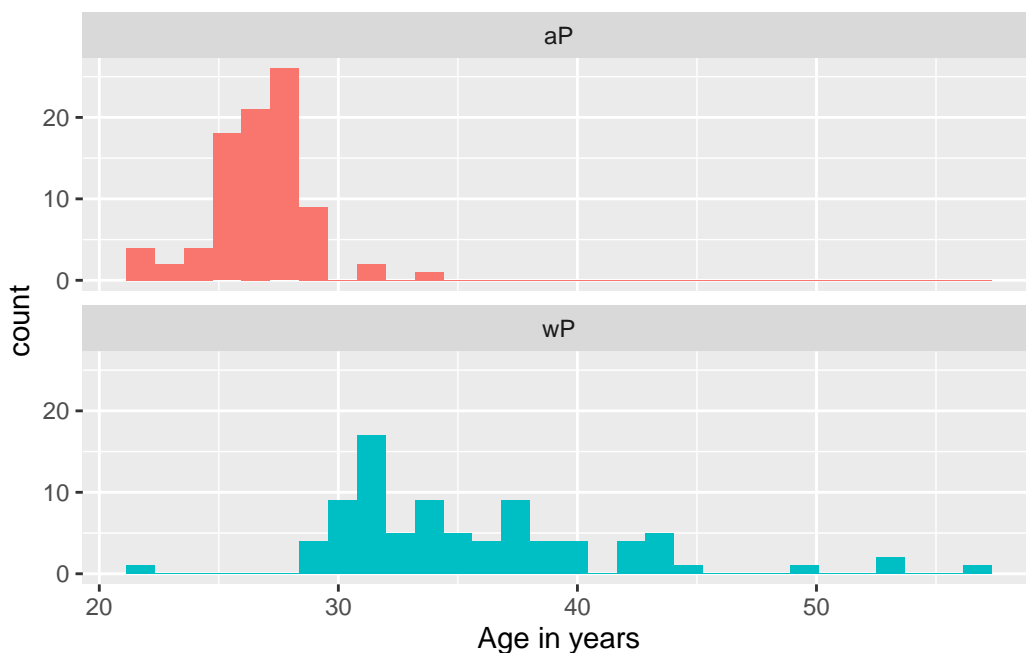
```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

```
[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) +
  xlab("Age in years")
```

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



Based on the graphs alone, the two groups look very different in their ages, however we can run a statistical significant report to confirm the difference.



```
x <- t.test(time_length( wp$age, "years" ),
            time_length( ap$age, "years" ))

x$p.value
```

```
[1] 2.372101e-23
```

## Joining multiple tables

Reading the specimen and ab\_titer tables and storing as dataframes.

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

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:

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

Joining with `by = join\_by(subject\_id)`

```
dim(meta)
```

```
[1] 1503    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	14210 days	38.90486
2	14210 days	38.90486
3	14210 days	38.90486
4	14210 days	38.90486
5	14210 days	38.90486
6	14210 days	38.90486

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.

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

Joining with `by = join\_by(specimen\_id)`

```
dim(abdata)
```

```
[1] 52576    22
```

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	5389	10117	10124	10124	10124

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

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

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

The difference in these values is that they vary a lot according to year and the most recent dataset has the lowest number of rows.

#### 4. Examine IgG Ab tier levels

```
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	actual_day_relative_to_boost
1	IU/ML	0.530000	1	-3
2	IU/ML	6.205949	1	-3
3	IU/ML	4.679535	1	-3
4	IU/ML	0.530000	3	-3
5	IU/ML	6.205949	3	-3
6	IU/ML	4.679535	3	-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		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

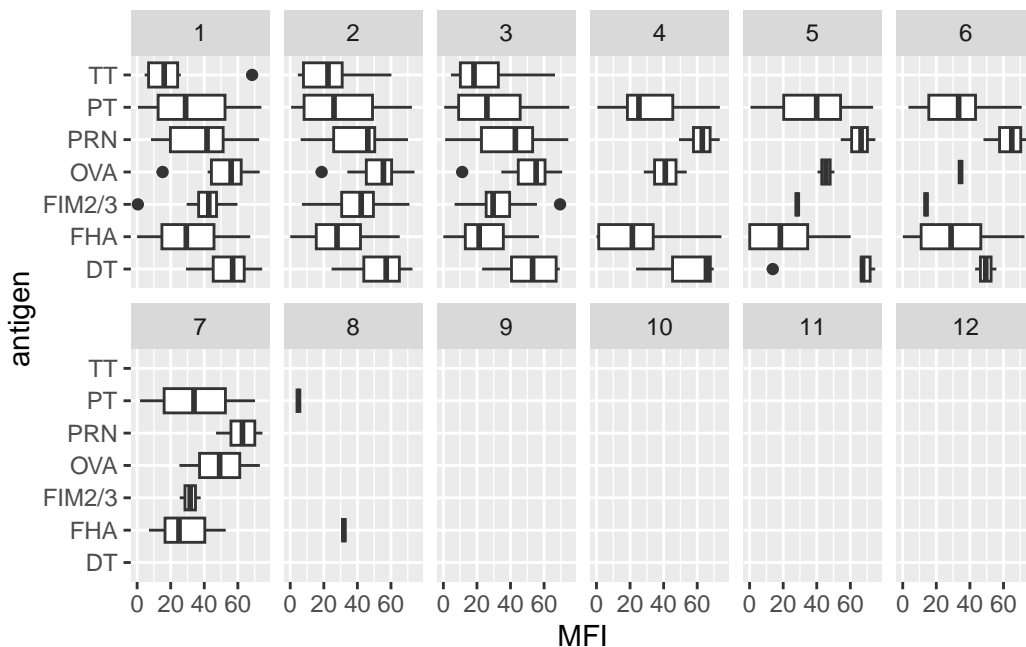
  

	age	age_years
1	14210 days	38.90486
2	14210 days	38.90486
3	14210 days	38.90486
4	15306 days	41.90554
5	15306 days	41.90554
6	15306 days	41.90554

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

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

Warning: Removed 4475 rows containing non-finite outside the scale range (`stat\_boxplot()`).

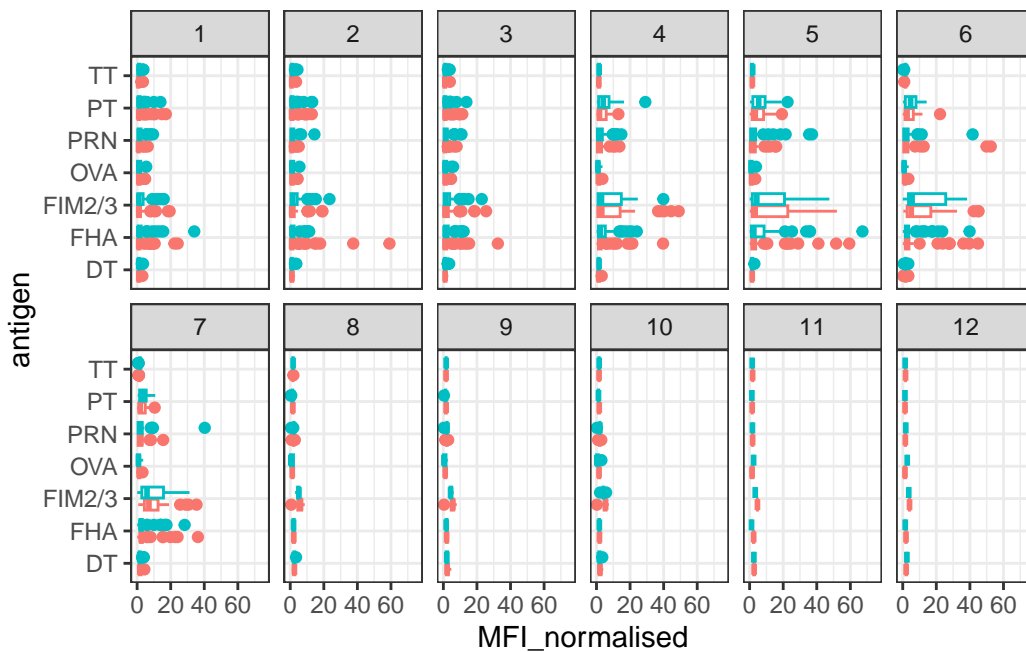


Q14. What antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others?

PT, PRN, FIM2/3, and FHA show differences in the level of IgG antibody titers. It is because they are administered widely, they can have bigger range of differences compared to the other vaccines.

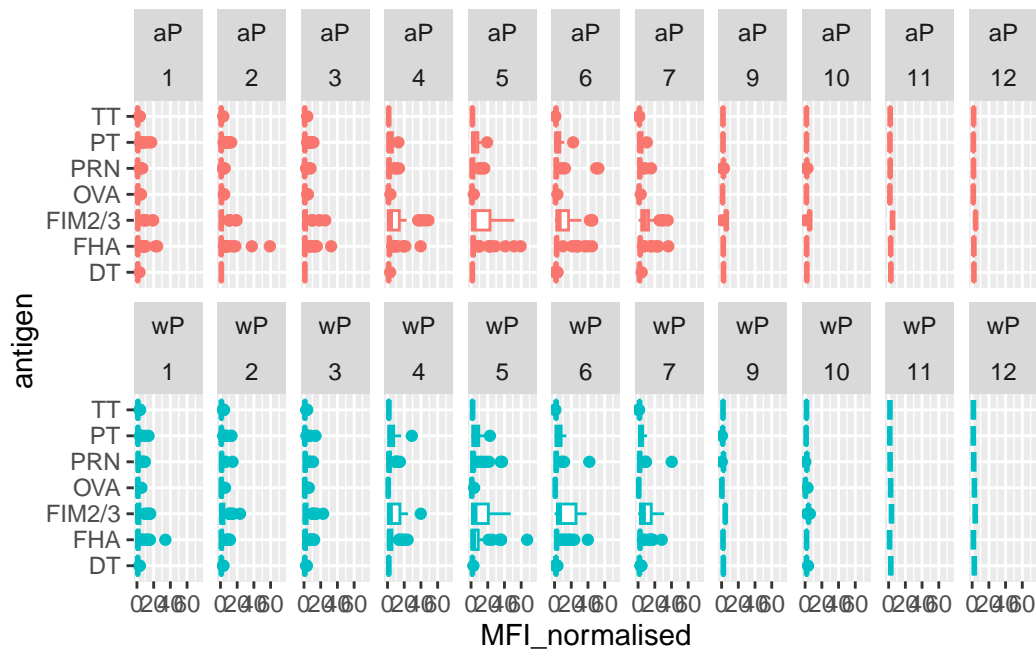
```
ggplot(igg) +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  xlim(0,75) +
  theme_bw()
```

Warning: Removed 5 rows containing non-finite outside the scale range (`stat\_boxplot()`).



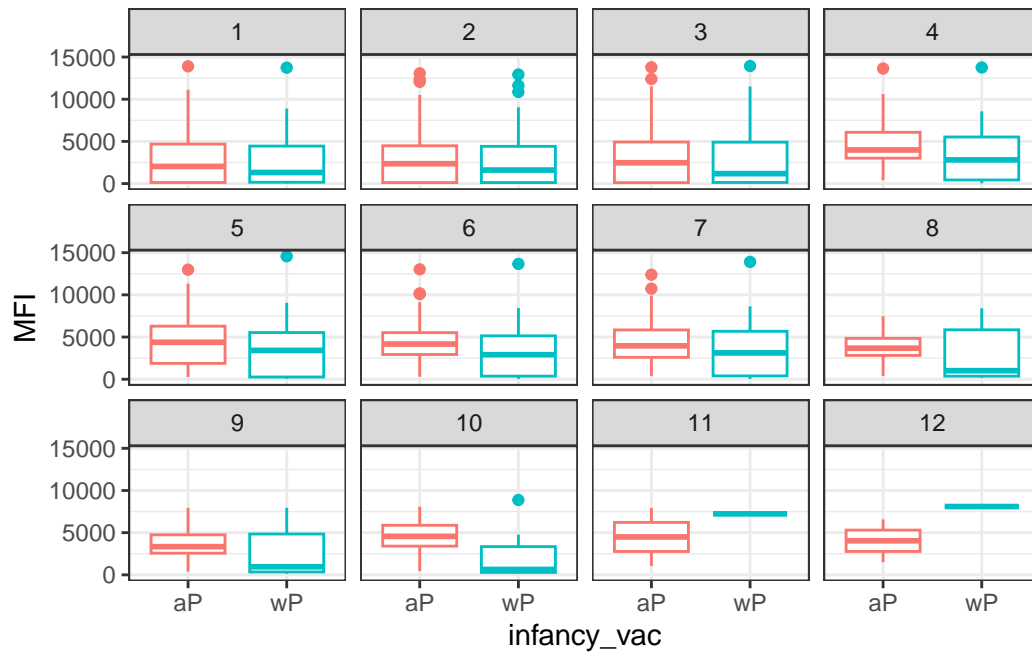
```
igg %>% filter(visit != 8) %>%
ggplot() +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  xlim(0,75) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```

Warning: Removed 5 rows containing non-finite outside the scale range  
(`stat\_boxplot()`).

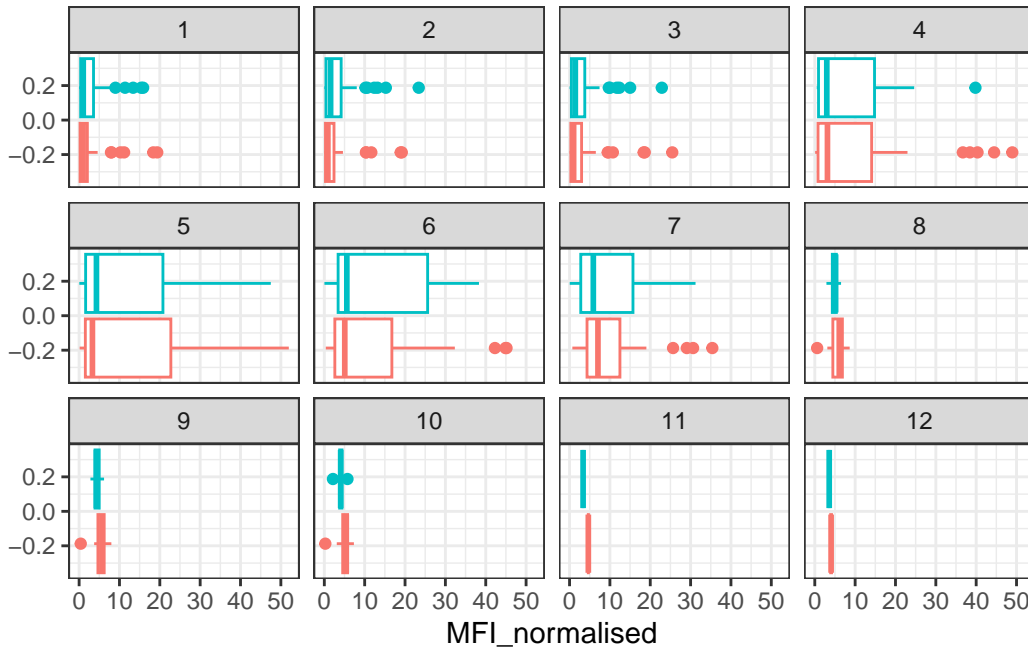


Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a “control” antigen (“OVA”, that is not in our vaccines) and a clear antigen of interest (“PT”, Pertussis Toxin, one of the key virulence factors produced by the bacterium *B. pertussis*).

```
filter(igg, antigen=="OVA") %>%
  ggplot() +
  aes(x = infancy_vac, y = MFI, 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()
```



Q16. What do you notice about these two antigens time courses and the PT data in particular?

The PT levels rise over time while the OVA levels don't show a consistent increase. This is for both the wP and aP subjects of PT.

Q17. Do you see any clear different in aP vs. wP responses?

I don't see any clear trends or differences between the aP and wP responses in OVA antigen, however for the PT antigen, the wP group shows slightly higher IgG titers than the aP group.

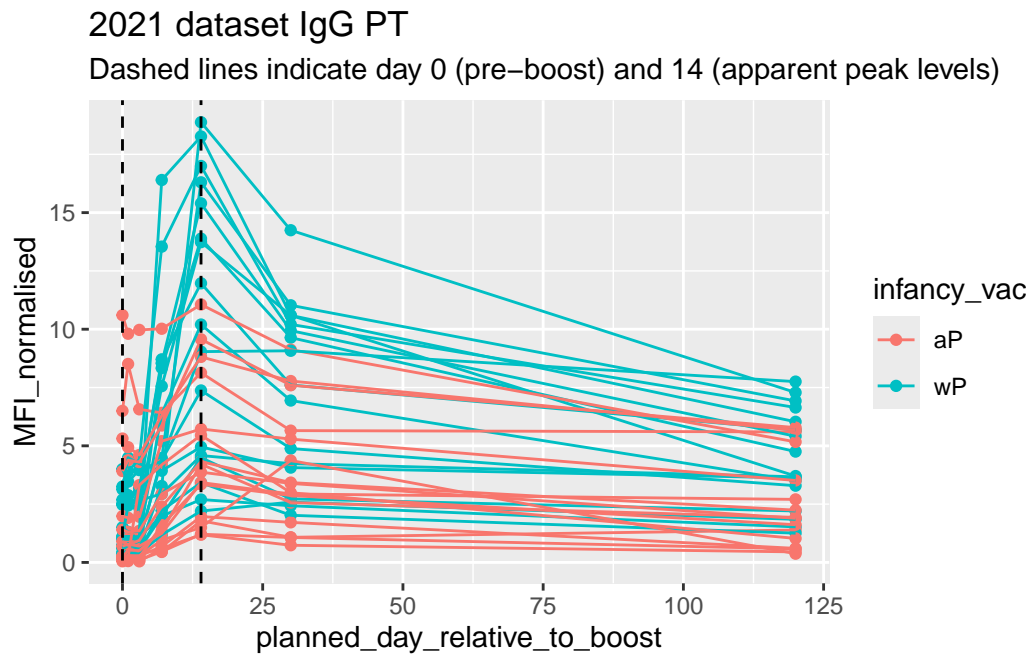
Looking at the 2021 dataset IgG PT antigen levels-time-course:

```
abdata.21 <- abdata %>% filter(dataset == "2021_dataset")

abdata.21 %>%
  filter(isotype == "IgG", antigen == "PT") %>%
  ggplot() +
    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(title="2021 dataset IgG PT",
      subtitle = "Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)")
```



Q18. Does this trend look similar for the 2020 dataset?

This trend looks similar to the 2020 dataset because in this too, there is a greater peak for the wP than the aP group.

###. 5. Obtaining CMI-PB RNASeq data

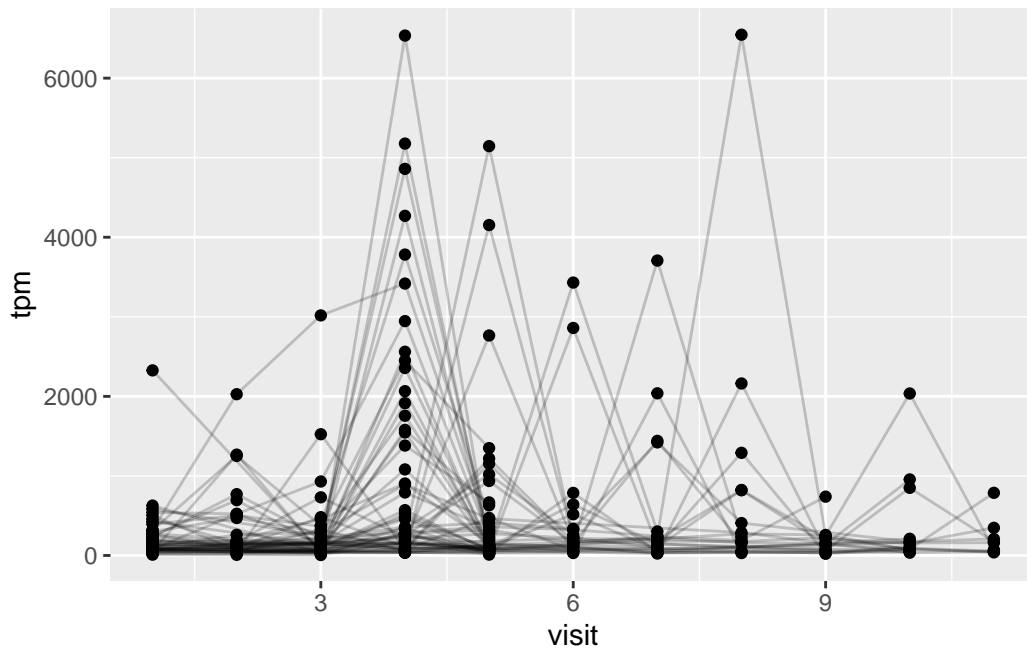
```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSOG00000211896.7"
rna <- read_json(url, simplifyVector = TRUE)
```

```
#meta <- inner_join(specimen, subject)
ssrna <- inner_join(rna, meta)
```

Joining with `by = join\_by(specimen\_id)`

Q19. Make a plot of the time course of gene expression for IGHG1 gene (i.e. a plot of visit vs. tpm).

```
ggplot(ssrna) +
  aes(x = visit, y = tpm, group=subject_id) +
  geom_point() +
  geom_line(alpha=0.2)
```



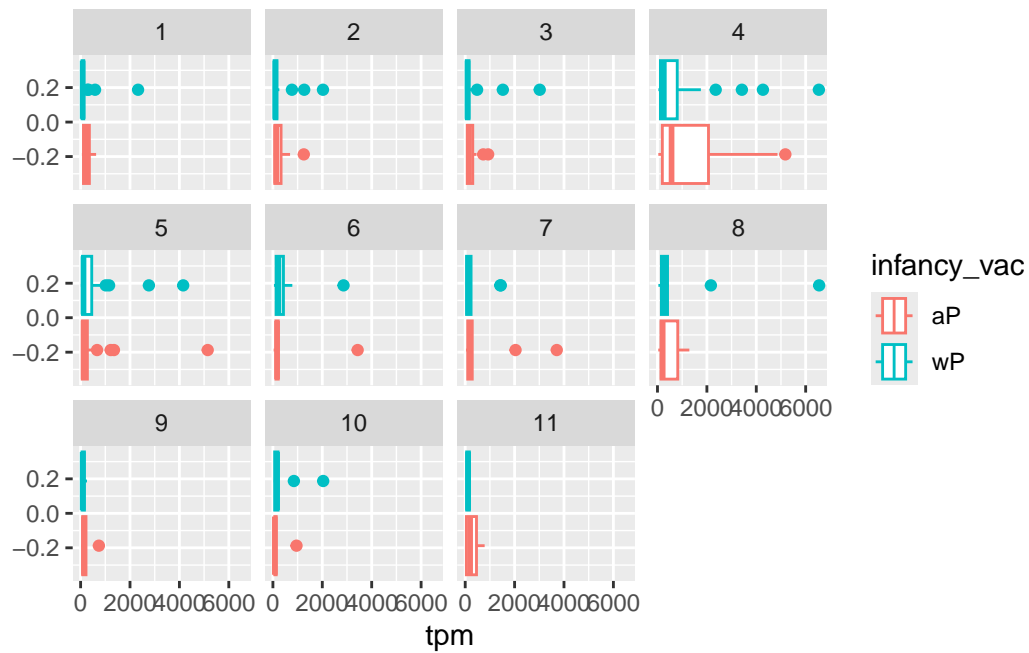
Q20.: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

The gene's maximum level happens at visit 4 (of over 6000 tpm).

Q21. Does this pattern in time match the trend of antibody titer data? If not, why not?

This pattern does match the trend of antibody titer data because both show that after the boost, the data peaks.

```
ggplot(ssrna) +
  aes(tpm, col=infancy_vac) +
  geom_boxplot() +
  facet_wrap(vars(visit))
```



```
ssrna %>%
  filter(visit==4) %>%
  ggplot() +
    aes(tpm, col=infancy_vac) + geom_density() +
    geom_rug()
```

