

# Lab 19

Karen Guerrero

```
##install.packages("datapasta")
```

## Is Pertussis on the rise

Scrape the CDC data on Pertussis cases per year in the US from their website here:

<https://www.cdc.gov/pertussis/surv-reporting/cases-by-year.html>

```
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),
  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)
```

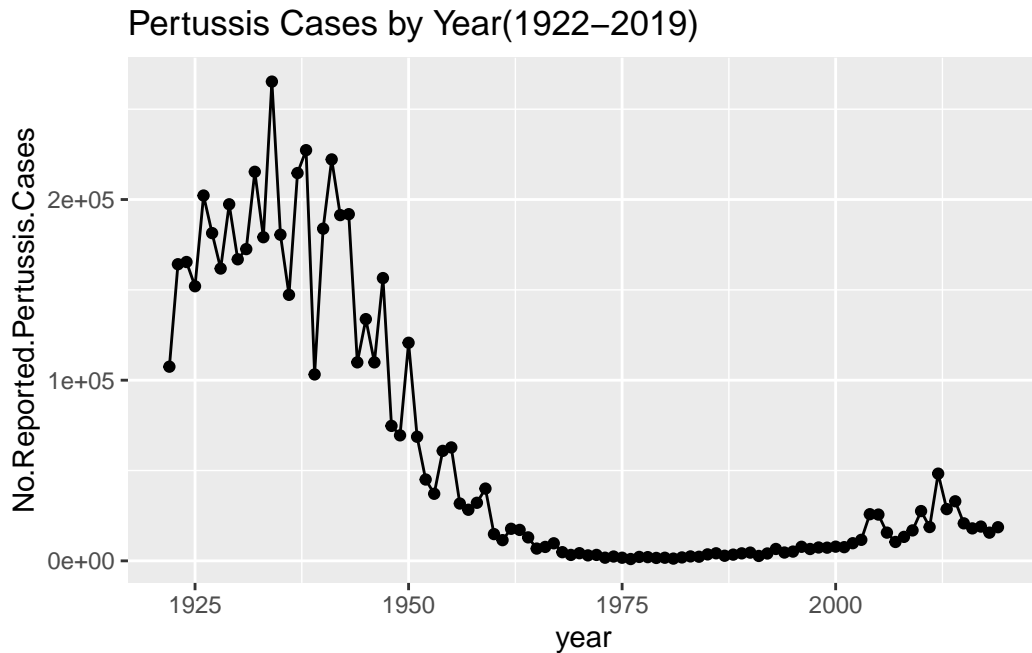
```
)
```

```
View(cdc)
```

```
library(ggplot2)
```

##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.

```
ggplot(cdc) +  
  aes(Year, No..Reported.Pertussis.Cases) +  
  geom_point() +  
  geom_line() +  
  labs(x="year", y="No.Reported.Pertussis.Cases", title="Pertussis Cases by Year(1922-2019)
```

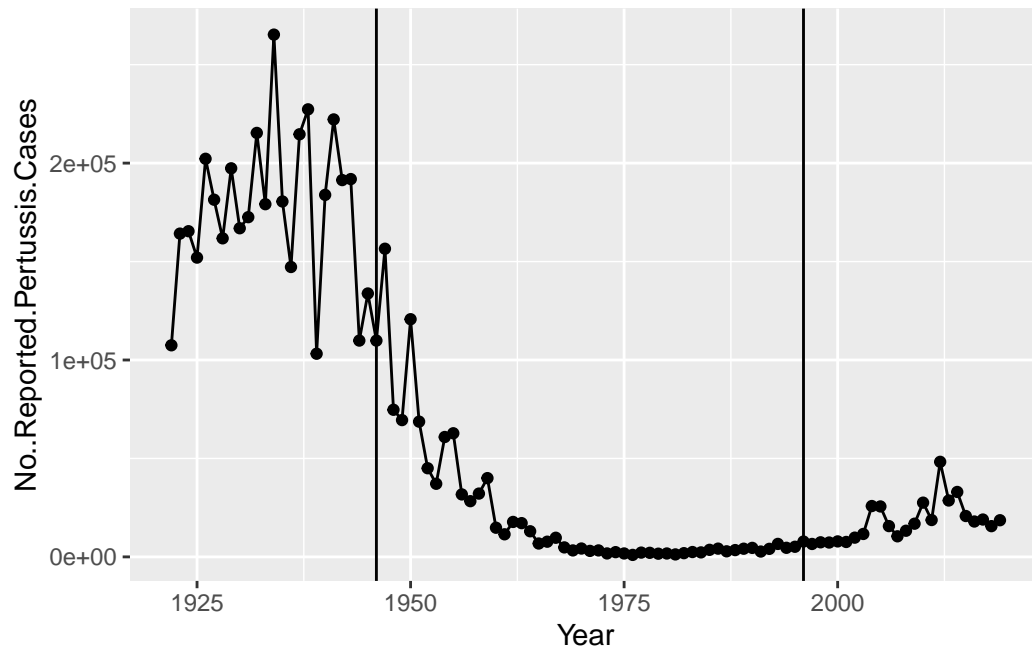


```
baseplot <- ggplot(cdc) +
  aes(Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +
  labs(x="year", y="No.Reported.Pertussis.Cases", title="Pertussis Cases by Year(1922–2019)
```

##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?

I noticed that there was a drastic decrease of number of reported Pertussis cases between 1946 and 1996. The number of reported Pertussis cases were very low by year ~1960 and stayed low until ~2002. After ~2002 there was a small increase of reported cases.

```
ggplot(cdc) +
  aes(Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() + geom_vline(xintercept=1996) + geom_vline(xintercept=1946)
```



```
labs(x="year", y="No.Reported.Pertussis.Cases", title="Pertussis Cases by Year(1922-2019)
```

```
$x
[1] "year"

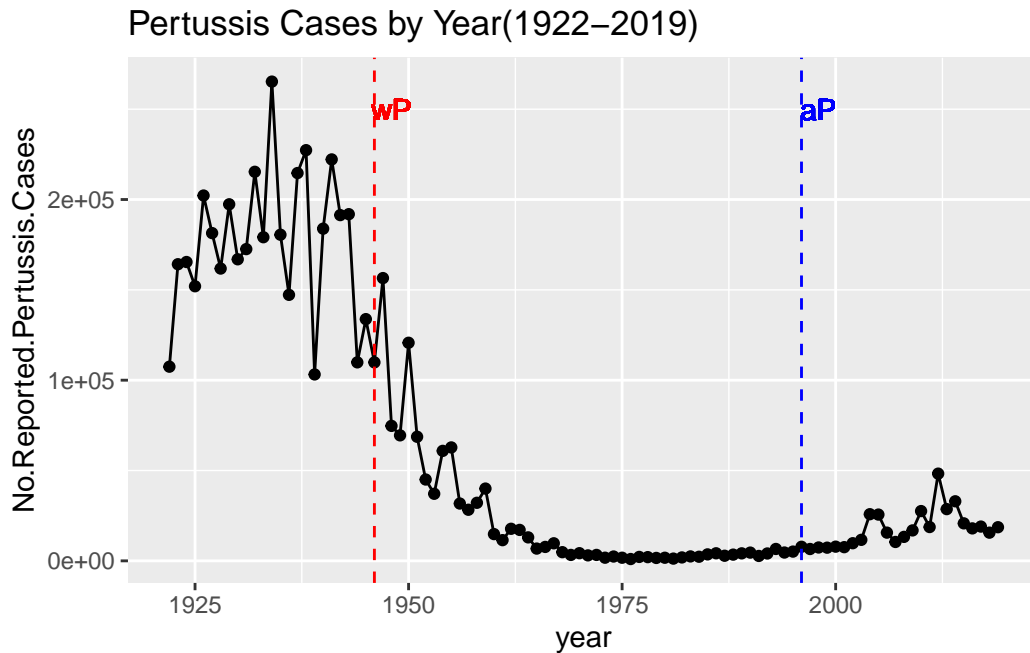
$y
[1] "No.Reported.Pertussis.Cases"

$title
[1] "Pertussis Cases by Year(1922-2019)"

attr("class")
[1] "labels"
```

There are two main Pertussis vaccine types, so-called whole-cell Pertusis (wP) and acellular Pertussis vaccines (aP). The older wP was introduced in 1946 and was highly effective. The switch to aP was in 1996.

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



##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 aP vaccine the number of reported Pertussis cases began to slightly increase. I think introducing a new type of Pertussis vaccine created a fear towards the new type of vaccine causing the population to not get vaccinated anymore due to fear.

## The CMI-PB database

The CMI-RB project and associated database is making data on the immune response to Pertussis boost available to the scientific community.

We will see what type of data CMI-PB make available here in a moment.

First we need to be able to read data from CMI-PB. It, like many large advanced databases, returns JSON format.

```
# 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?

There are 47 aP infancy vaccinated subjects. There are 49 wP infancy vaccinated subjects.

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

```
aP wP
47 49
```

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

There are 66 female subjects and 30 male subjects in the dataset.

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

```
Female  Male
   66    30
```

```
#View(subject)
```

##Q6. What is the breakdown of race and biological sex (e.g. number of Asian females, White males etc...)?

American Indian/Alaska Native: 0 Females, 1 Male Asian: 18 Females, 9 Males Black or African American: 2 Females, 0 Male More Than One Race: 8 Females, 2 Males Native Hawaiian or Other Pacific Islander: 1 Female, 1 Male Unknown or Not Reported: 10 Females, 4 Males White: 27 Females, 13 Males

```
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

```
library(lubridate)
```

Loading required package: timechange

Attaching package: 'lubridate'

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

date, intersect, setdiff, union

##What is today's date?

```
today()
```

```
[1] "2022-12-02"
```

##How many days have passed since new year 2000?

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

Time difference of 8371 days

##What is this in years?

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

```
[1] 22.91855
```

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

```
round( summary(ap$age))
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
23	25	26	25	26	27

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

```
round( summary(wp$age))
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
28	32	35	36	40	55

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



### Welch Two Sample t-test

```
data: ap$age and wp$age
t = -12.092, df = 51.082, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.644857 -9.044045
sample estimates:
mean of x mean of y
 25.23634  36.08079
```

##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?

The average age of wP individuals is 36.

The average age of aP individuals is 25.

Based off the p-value (2.2e-16) obtained from a T-test, it demonstrates that they are significantly different.

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

The ages of all individuals at time of boost are 30, 51, 33, 28, 25 and 28.

```
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
```

What is the average age across all subjects?

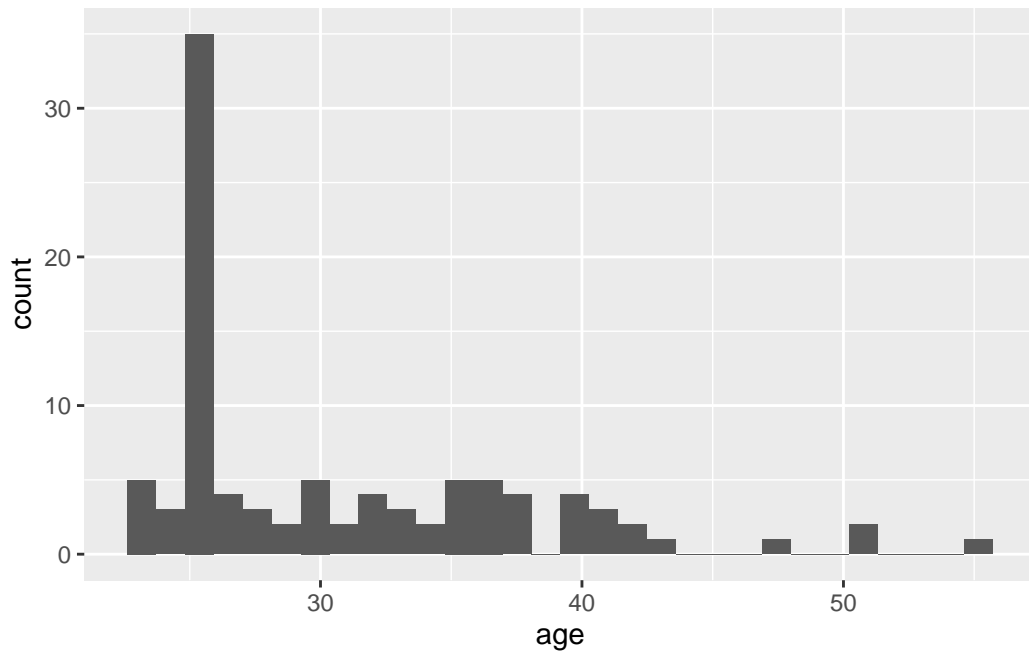
```
summary(subject$age)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
22.92	25.92	27.92	30.77	34.92	54.92

Histogram

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

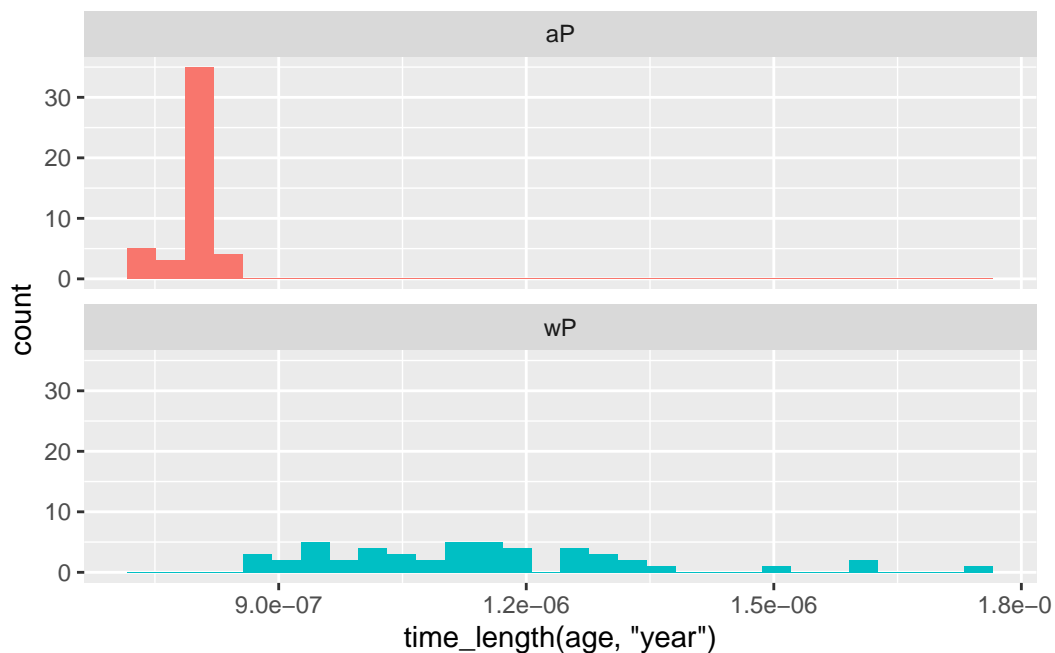
``stat_bin()`` using ``bins = 30``. Pick better value with ``binwidth``.



And faceted by wP and aP

```
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``.



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

```
x$p.value
```

```
[1] 1.316045e-16
```

##Q9. With the help of a faceted boxplot or histogram, do you think these two groups are significantly different?

Yes, based off the p-value (1.316045e-16) both groups are significantly different.

```
specimen <- read_json("https://www.cmi-pb.org/api/specimen", simplifyVector = TRUE)
titer <- read_json("https://www.cmi-pb.org/api/ab_titer", 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 ‘specimen’ tables to have all the metadata we need for later analysis. We can use the dplyr ‘\*\_join()’ functions for this task.

##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:

```
library(dplyr)
meta <- inner_join(specimen, subject)
```

Joining, by = "subject\_id"

```
dim(meta)
```

```
[1] 729 14
```

```
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 link a.k.a join this ab titer data with our meta data to have all the information we need to do some actual analysis like comparing Ab levels at different times after boost, comparing ap vs wP responses over time etc.

##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, by = "specimen\_id"

```
dim(abdata)
```

```
[1] 32675    21
```

##Q11. How many specimens (i.e. entries in abdata) do we have for each isotype?

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

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

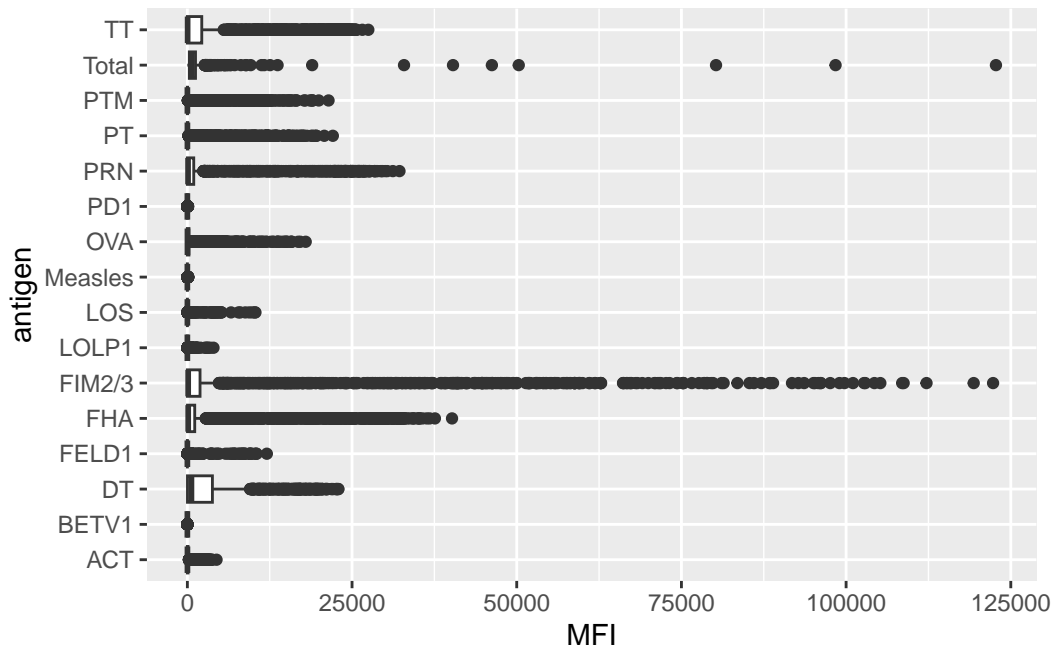
```

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

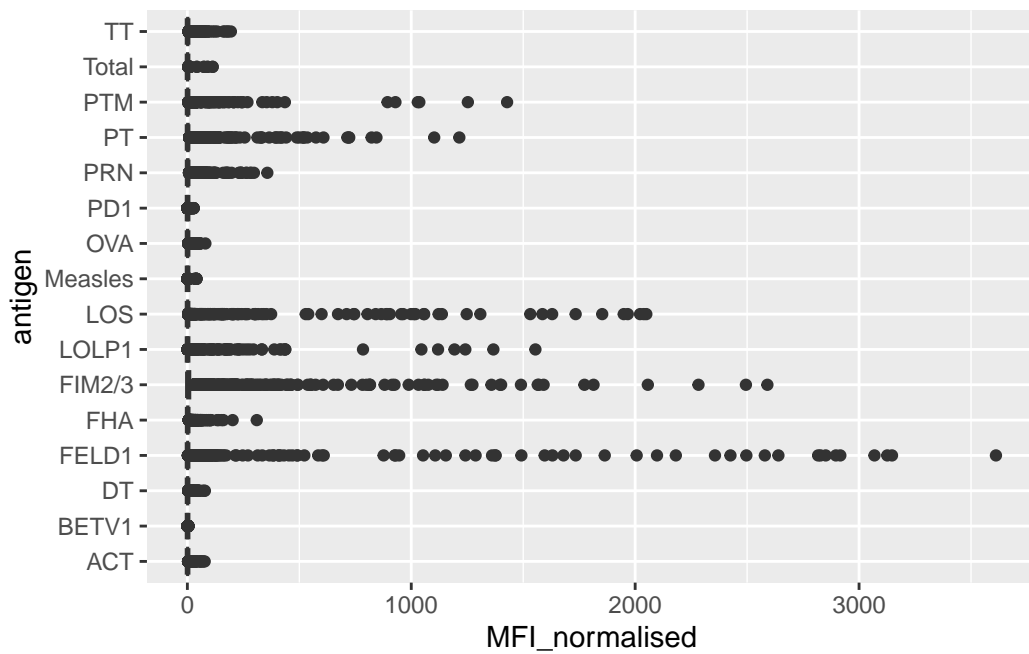
```

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

Warning: Removed 1 rows containing non-finite values (`stat\_boxplot()`).



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



##Q12. What do you notice about the number of visit 8 specimens compared to other visits?

I notice that the number of visit 8 specimens is much lower compared to the other visits.

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

```

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

```

```
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

```

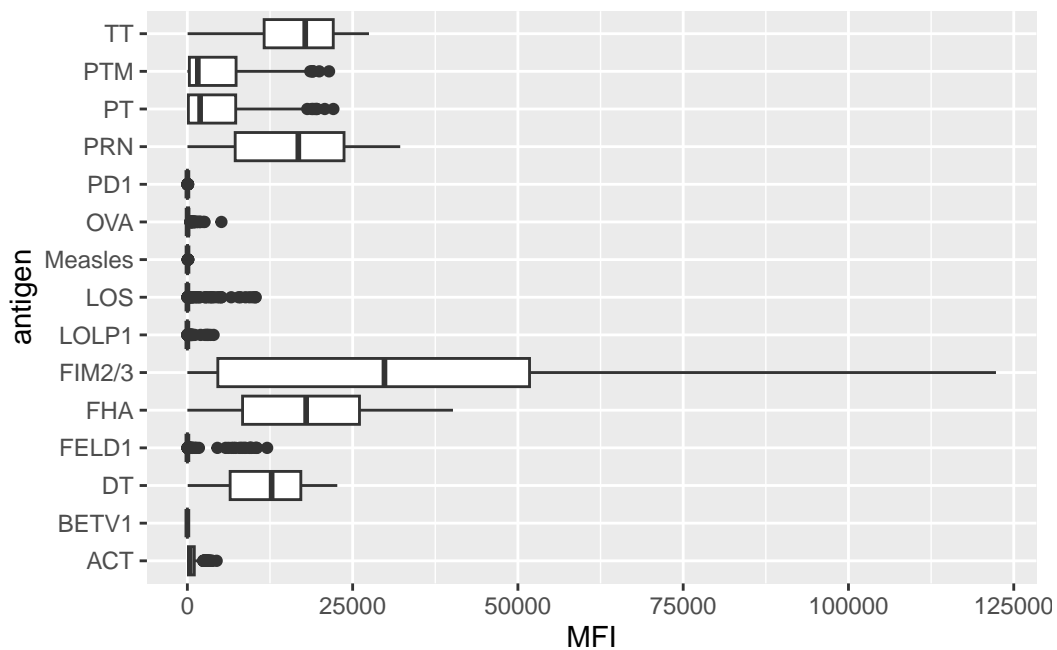
6 Not Hispanic or Latino White    1986-01-01    2016-09-12 2020_dataset
  age
1 36.91718
2 36.91718
3 36.91718
4 36.91718
5 36.91718
6 36.91718

```

```

ggplot(ig1) +
  aes(MFI, antigen) +
  geom_boxplot()

```



We can break this up by visit (a quick proxy for time since boost).

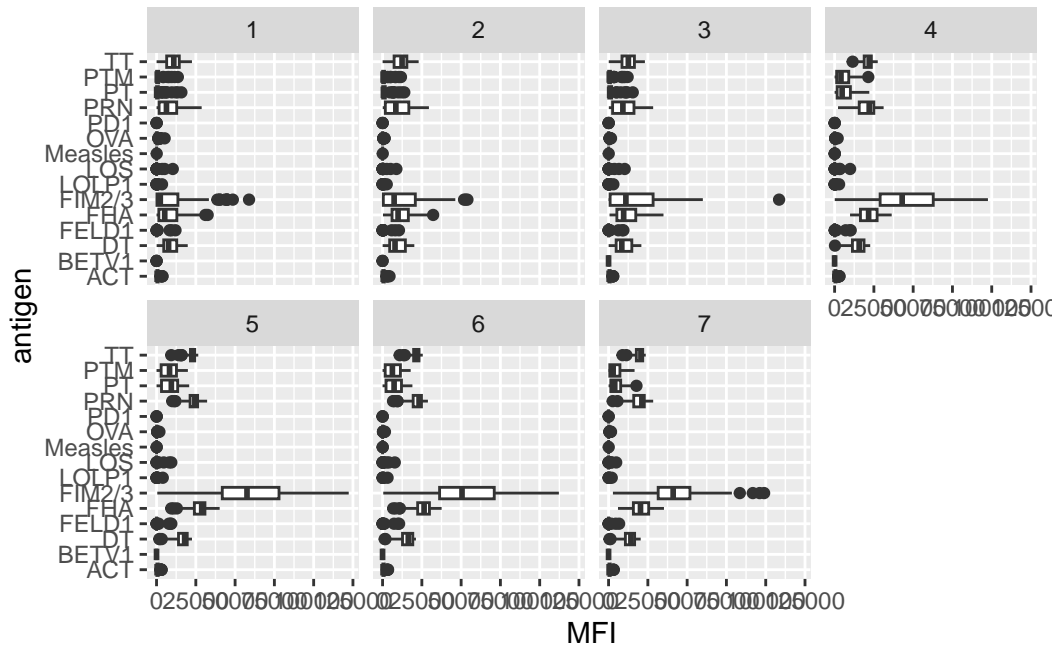
##Q13. Complete the following code to make a summary boxplot of Ab titer levels for all antigens:

```

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

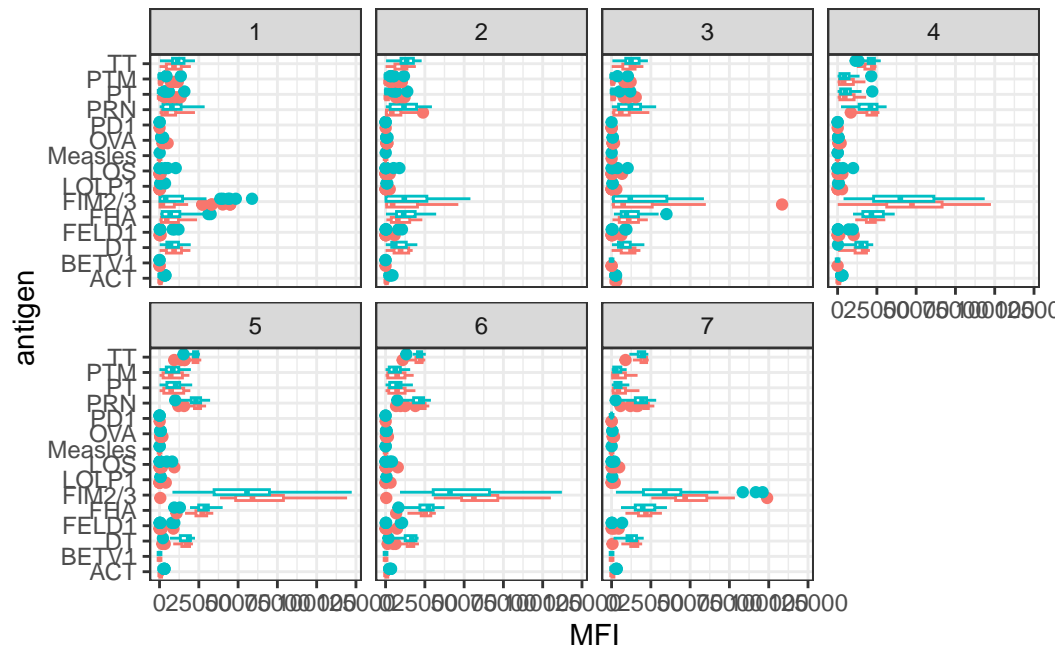
```



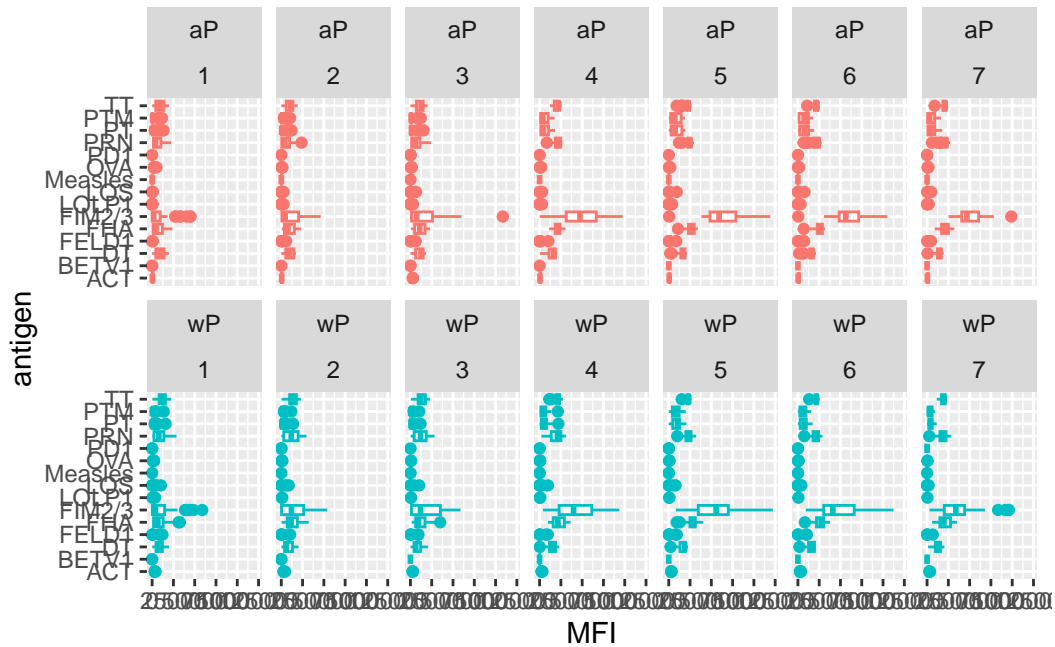


##Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others? Antigen FIM2/3 shows the greatest difference in the level of IgG1 antibody titers recognizing them over time. Antigen FIM2/3 has the greatest difference over time because it might have been used for the development of the vaccine, which would affect the level of IgG1 antibody titers recognizing them over time.

```
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)
```



##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 (“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 = TRUE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```

