

Class18

Sabrina Koldinger (A16368238)

Class 18

Lets examine and explore Pertusis cases in the US. URL: <https://www.cdc.gov/pertussis/surv-reporting/cases-by-year.html>

Use datapasta to scrape this data from the website into R:

```
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),  
  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  Cases
1 1922 107473
2 1923 164191
3 1924 165418
4 1925 152003
5 1926 202210
6 1927 181411

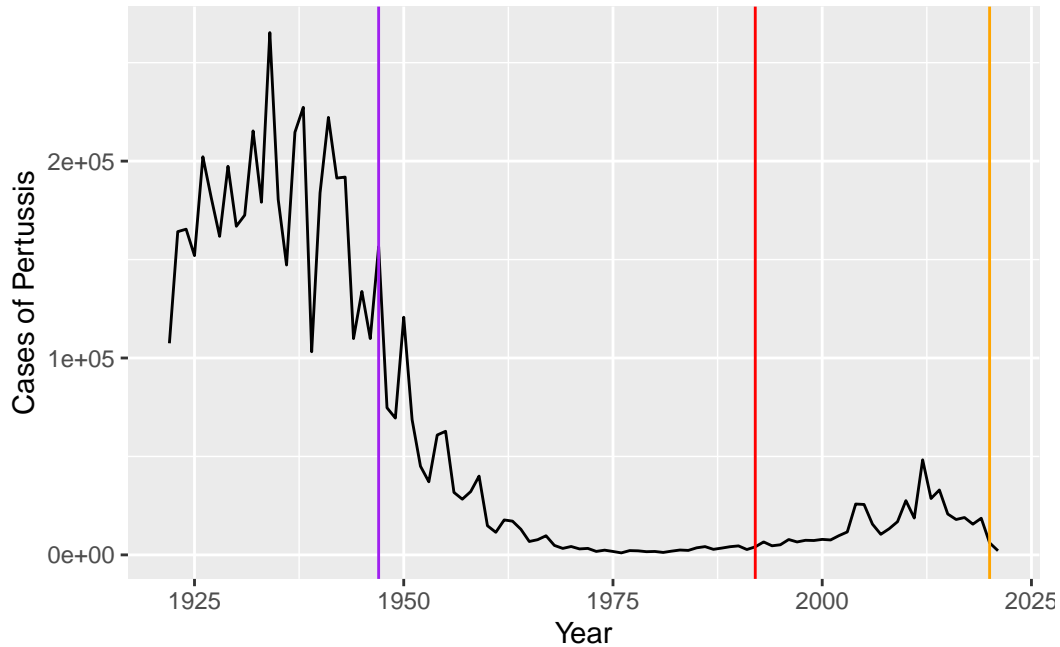
```

Plot it as cases per year

```

library(ggplot2)
ggplot(cdc) + aes(x=Year,y=Cases) + geom_line() + geom_vline(xintercept =1947, col= "purple") +
  geom_vline(xintercept =1992, col= "red")+
  geom_vline(xintercept =2020, col= "orange")+
  ylab("Cases of Pertussis")

```



Import data from CMI-PB project. This database uses API to return JSON format data. We will use `jsonlite`.

```
library(jsonlite)

subject=read_json("http://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

How many people have the wP vs aP (acellular vaccine) in the dataset?

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

```
aP wP  
60 58
```

60 with aP and 58 with wP

How many male or female are in the study?

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

```
Female    Male  
    79     39
```

79 women and 39 men

Their races? both combined?

```
table(subject$race)
```

```
        American Indian/Alaska Native  
                1  
                Asian  
                32  
        Black or African American  
                2  
        More Than One Race  
                11  
Native Hawaiian or Other Pacific Islander  
                2  
        Unknown or Not Reported  
                15  
                White  
                55
```

```
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

ages/ dates

We can use the lubridate package to ease the pain of dealing with dates.

```
library(lubridate)
```

Attaching package: 'lubridate'

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

date, intersect, setdiff, union

```
today()
```

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

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

Time difference of 8832 days

```
today() - ymd("2002-07-01")
```

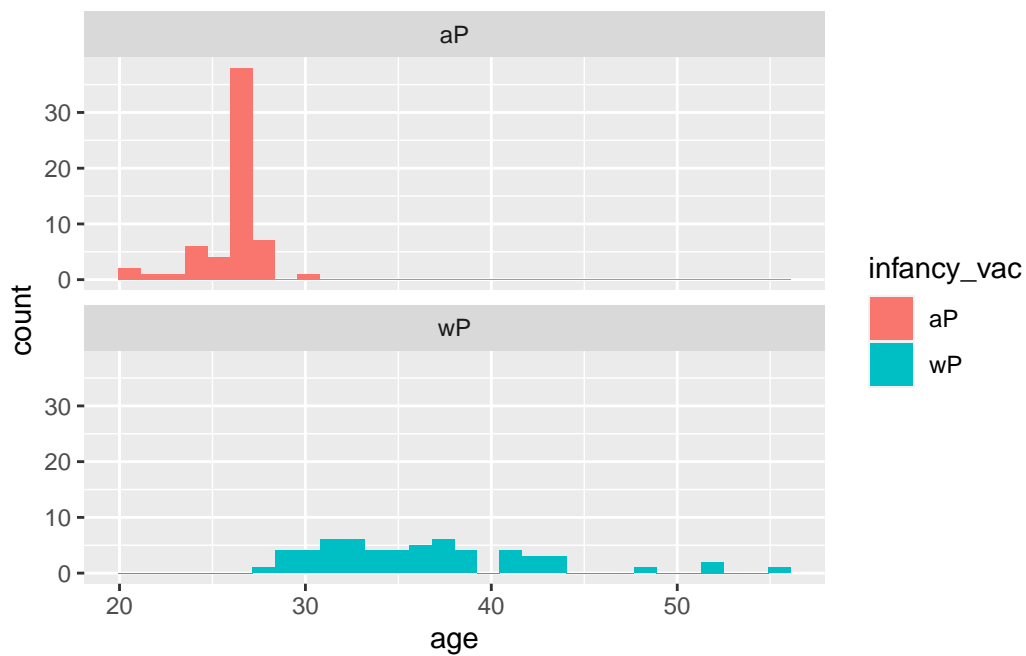
Time difference of 7920 days

What is the age of everyone on our dataset?

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

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

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



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

Use **join** to combine the specimen and subject tables to make a meta table with the metadata. We use the **dplyr** functions for this.

```
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	1

2	1986-01-01	2016-09-12	2020_dataset	38	2
3	1986-01-01	2016-09-12	2020_dataset	38	3
4	1986-01-01	2016-09-12	2020_dataset	38	4
5	1986-01-01	2016-09-12	2020_dataset	38	5
6	1986-01-01	2016-09-12	2020_dataset	38	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

Read in other data from CMI-PB

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

Combining meta and ab_titer with inner_join()


```
alldata= inner_join(meta, ab_titer)
```

Joining with `by = join_by(specimen_id)`

```
head(alldata)
```

	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	1
2	1986-01-01	2016-09-12	2020_dataset	38	1
3	1986-01-01	2016-09-12	2020_dataset	38	1
4	1986-01-01	2016-09-12	2020_dataset	38	1
5	1986-01-01	2016-09-12	2020_dataset	38	1
6	1986-01-01	2016-09-12	2020_dataset	38	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

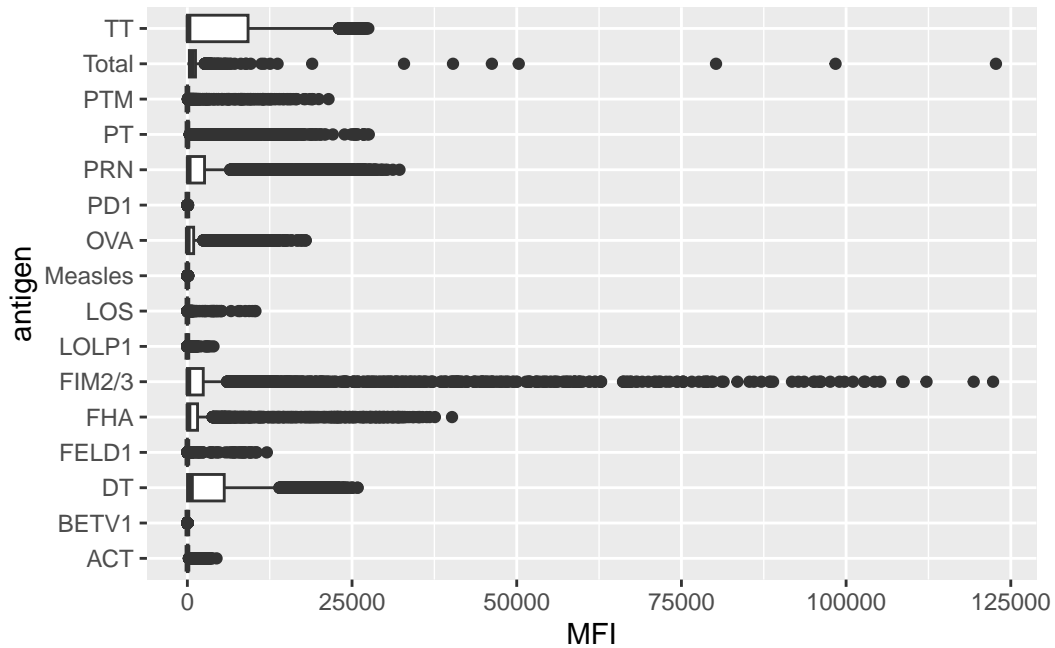
	visit	isotype	is_antigen_specific	antigen	MFI	MFI_normalised	unit
1	1	IgE	FALSE	Total	1110.21154	2.493425	UG/ML
2	1	IgE	FALSE	Total	2708.91616	2.493425	IU/ML
3	1	IgG	TRUE	PT	68.56614	3.736992	IU/ML
4	1	IgG	TRUE	PRN	332.12718	2.602350	IU/ML
5	1	IgG	TRUE	FHA	1887.12263	34.050956	IU/ML
6	1	IgE	TRUE	ACT	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

Plot of MFI vs. antigen

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

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

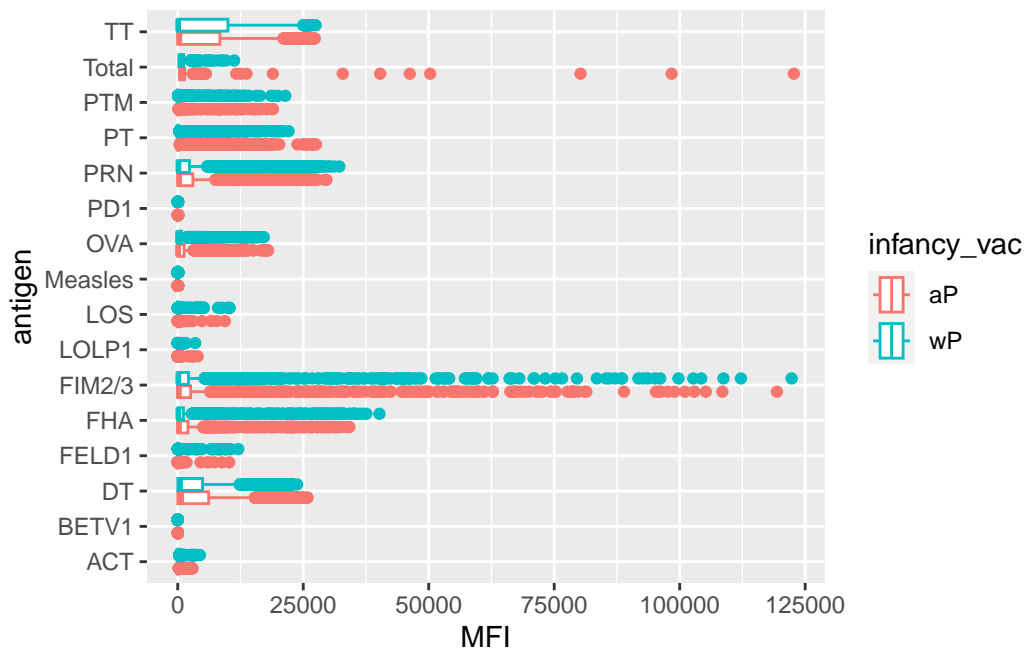


Why are certain antigens and not others very carable in their detected levels here?

Is there a difference between aP and wP?

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

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



Not a significant difference just by looking. There are some differences but hard to tell.

```
table(alldata$dataset)
```

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

Focus in on the 2021 dataset.

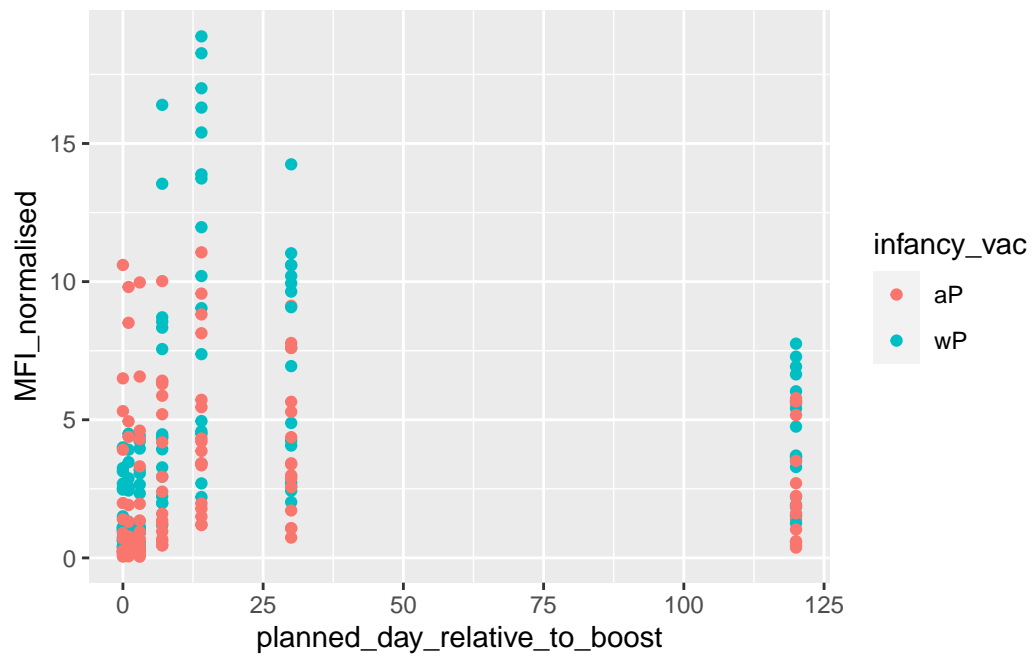
```
alldata.21=filter(alldata, dataset== "2021_dataset")
table(alldata.21$dataset)
```

```
2021_dataset
      8085
```

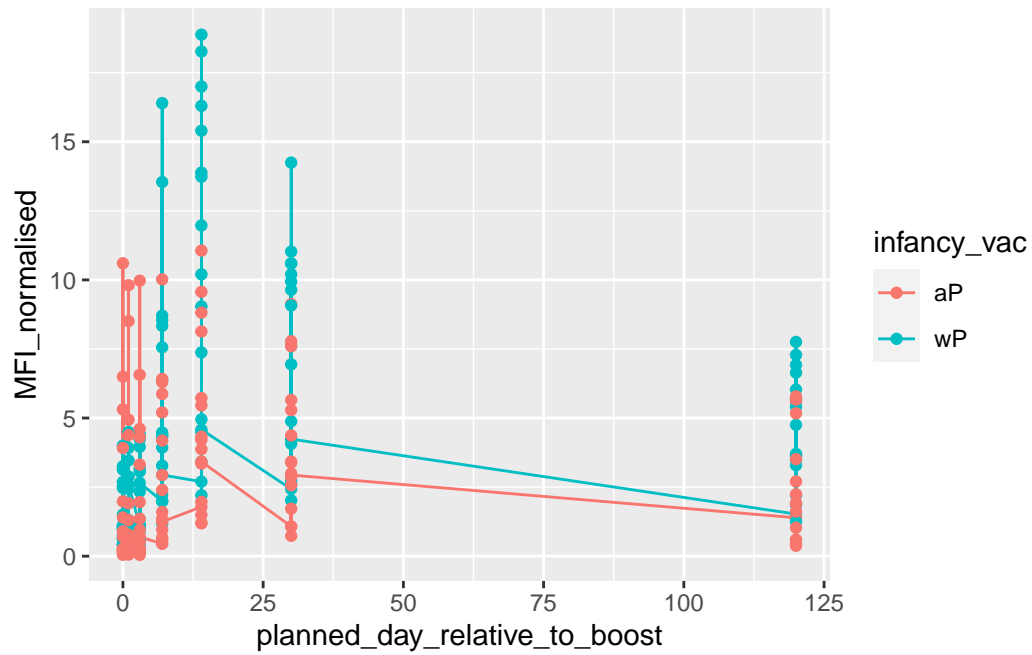
PT antigen IgG levels

```
pt.21= filter(alldata.21, isotype == "IgG", antigen == "PT")
```

```
ggplot(pt.21) + aes(x=planned_day_relative_to_boost, y=MFI_normalised, col=infancy_vac) + g
```



```
ggplot(pt.21) + aes(x=planned_day_relative_to_boost, y=MFI_normalised, col=infancy_vac) + g
```



```
ggplot(pt.21) + aes(x=planned_day_relative_to_boost, y=MFI_normalised, col=infancy_vac, gr
```

