

Lab18: Mini Project: Investigating Pertussis Resurgence

Duy An Le (PID: A16400411)

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.

Use datapasta package:

```
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),  
  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)
```

```
)
```

```
head(cdc)
```

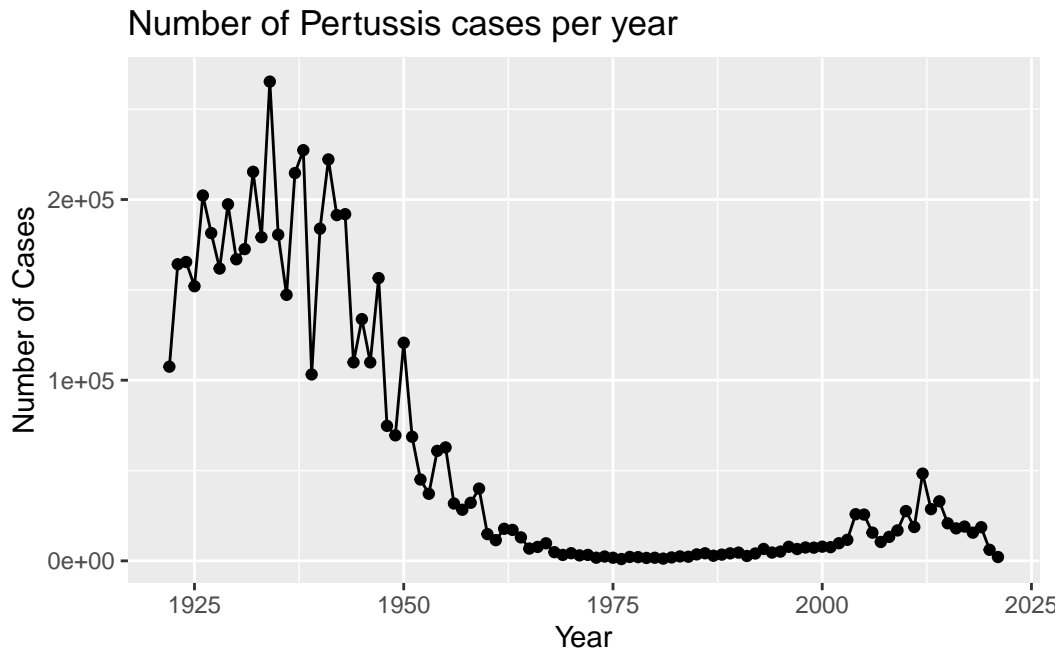
	Year	No..Reported.Pertussis.Cases
1	1922	107473
2	1923	164191
3	1924	165418
4	1925	152003
5	1926	202210
6	1927	181411

Make a plot with ggplot

```
library(ggplot2)
```

```
pertussis <- ggplot(cdc) + aes(x=Year, y=cdc[,2]) + geom_point() + labs(title="Number of P
```

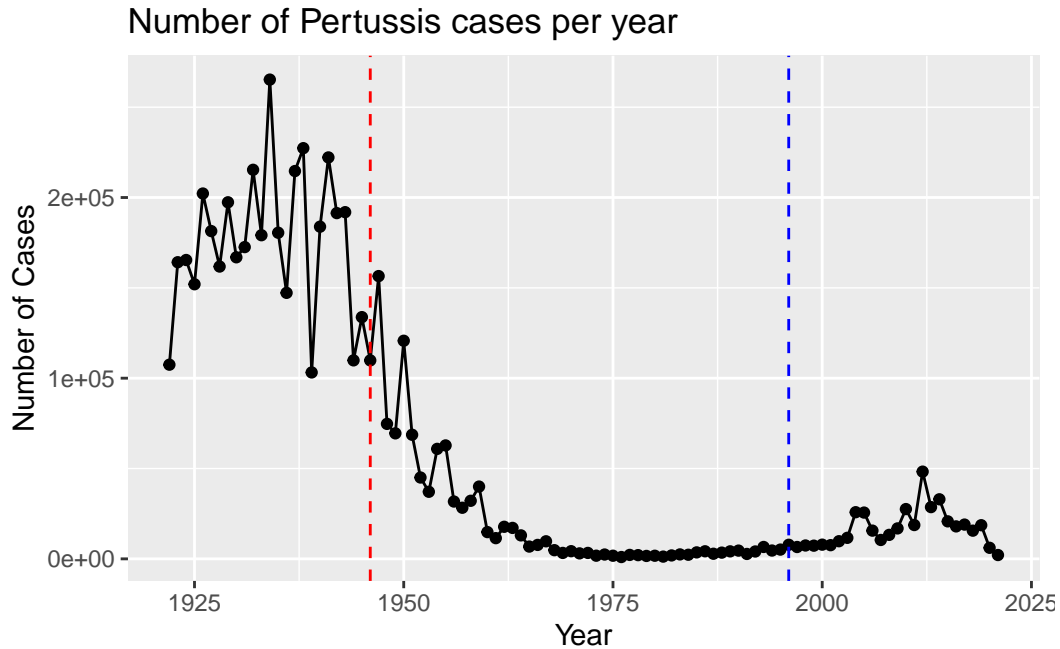
```
pertussis
```



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?

Add some lines to show addition of wP (1946) and aP (1996) vaccines.

```
pertussis + geom_vline(xintercept=c(1946,1996), colour=c("red","blue"), linetype="dashed")
```



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

After the addition of the wP vaccine, pertussis cases went down significantly over the next 25 or so years from 100,000 to near 0. However, after aP vaccine was added, there was a resurgence in cases in the year 2000 on. This could be due to the anti-vax campaign, causing parents to not vaccinate their children.

Use Jsonlite to read ongoing data from CMI-PB project

```
library(jsonlite)

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

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

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

```
aP wP
60 58
```

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

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

```
Female    Male
    79     39
```

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	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

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 lubridate package

```
library(lubridate)
```

Attaching package: 'lubridate'

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

```
date, intersect, setdiff, union
```

```
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
```

```
wp.ages <- time_length(today() - ymd((subject %>% filter(infancy_vac=="wP"))$year_of_birth))
summary(wp.ages)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
28.41	31.66	35.91	36.81	39.41	56.41

```
ap.ages <- time_length(today() - ymd((subject %>% filter(infancy_vac=="aP"))$year_of_birth))
summary(ap.ages)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
21.41	26.41	26.41	26.51	27.41	30.41

```
subject$age <- today() - ymd(subject$year_of_birth)
```

The average age of wP individuals is 37 and the average age of aP individuals is 27. On average, wP individuals are 10 years older than aP individuals.

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

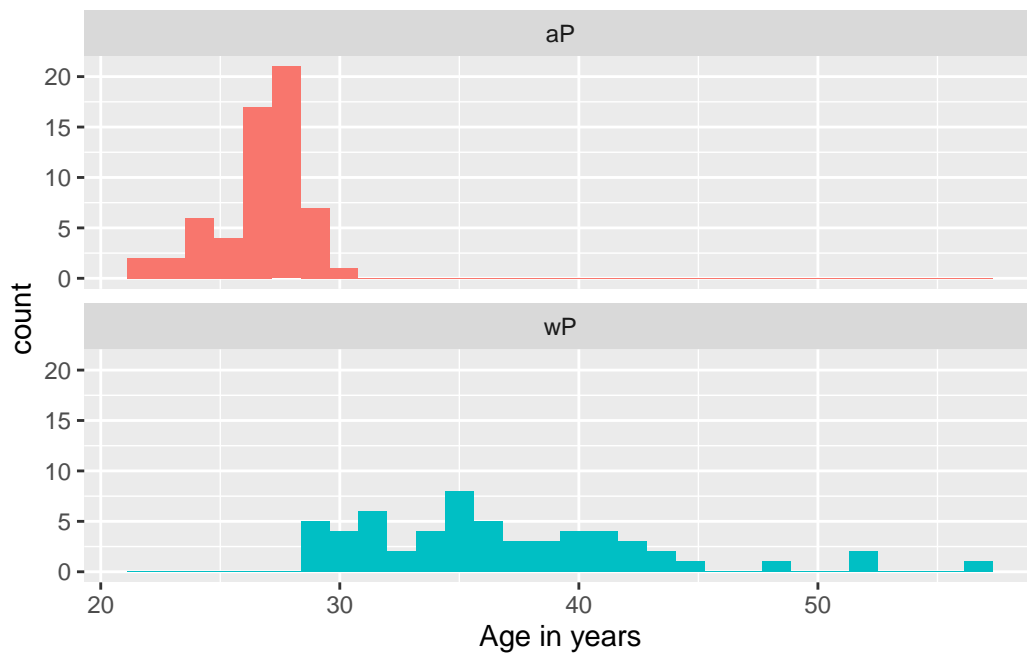
```
age.at.boost <- time_length((ymd(subject$date_of_boost) - ymd(subject$year_of_birth)), "year")
head(age.at.boost)
```

```
[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
```

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



Obtain titer and specimen values from CMI-PB API

```
specimen <- read_json("https://www.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

```
titer <- read_json("https://www.cmi-pb.org/api/plasma_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

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:

Join the databases using inner_join (common column between x,y used to link)

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

Joining with `by = join_by(subject_id)`

```
dim(meta)
```

```
[1] 939  14
```

```
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	14029 days	1
2	1986-01-01	2016-09-12	2020_dataset	14029 days	2
3	1986-01-01	2016-09-12	2020_dataset	14029 days	3
4	1986-01-01	2016-09-12	2020_dataset	14029 days	4
5	1986-01-01	2016-09-12	2020_dataset	14029 days	5
6	1986-01-01	2016-09-12	2020_dataset	14029 days	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

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] 46906    21
```

```
head(abdata)
```

	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	subject_id	infancy_vac	biological_sex
1	UG/ML	2.096133	1	wP	Female
2	IU/ML	29.170000	1	wP	Female
3	IU/ML	0.530000	1	wP	Female
4	IU/ML	6.205949	1	wP	Female
5	IU/ML	4.679535	1	wP	Female
6	IU/ML	2.816431	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	actual_day_relative_to_boost	planned_day_relative_to_boost
1	14029 days	-3	0
2	14029 days	-3	0
3	14029 days	-3	0
4	14029 days	-3	0

```

5 14029 days -3 0
6 14029 days -3 0
specimen_type visit
1 Blood 1
2 Blood 1
3 Blood 1
4 Blood 1
5 Blood 1
6 Blood 1

```

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 4255 8983 8990 8990 8990

```

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$antigen)
```

```

ACT BETV1 DT FELD1 FHA Fim2/3 FIM2/3 LOLP1 LOS Measles
1970 1970 4168 1970 4562 1043 3125 1970 1970 1970
OVA PD1 PRN PT PTM Total TT
4168 1970 4562 4562 1970 788 4168

```

Examining IgG antibody titer 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

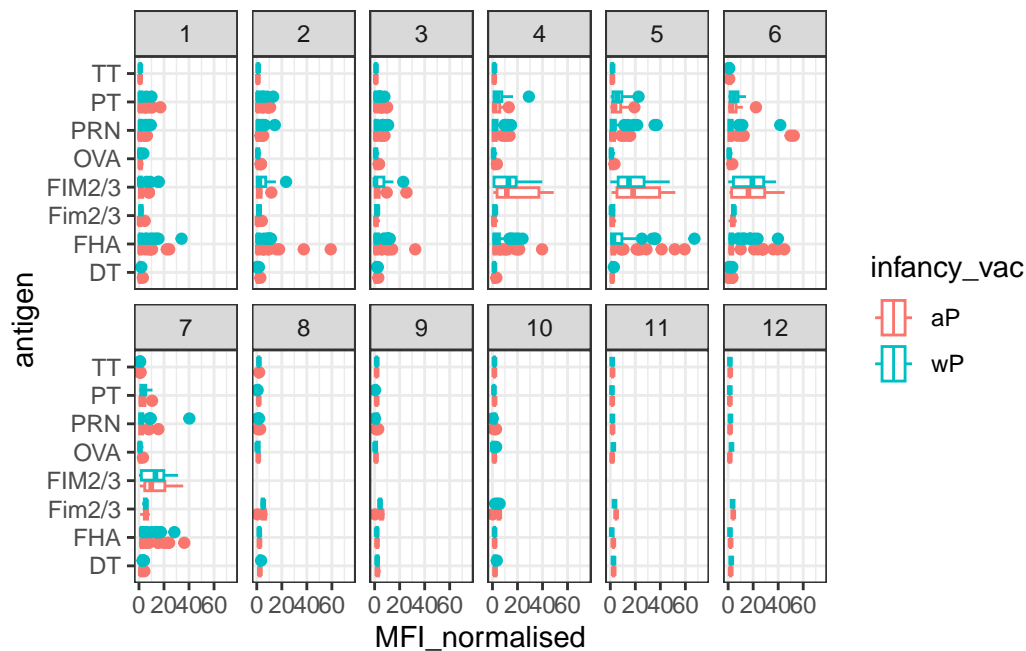
```

6	19	IgG	TRUE	FHA	60.76626	1.096457
	unit	lower_limit_of_detection	subject_id	infancy_vac	biological_sex	
1	IU/ML	0.530000	1	wP	Female	
2	IU/ML	6.205949	1	wP	Female	
3	IU/ML	4.679535	1	wP	Female	
4	IU/ML	0.530000	3	wP	Female	
5	IU/ML	6.205949	3	wP	Female	
6	IU/ML	4.679535	3	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	actual_day_relative_to_boost	planned_day_relative_to_boost			
1	14029	days	-3			0
2	14029	days	-3			0
3	14029	days	-3			0
4	15125	days	-3			0
5	15125	days	-3			0
6	15125	days	-3			0
	specimen_type	visit				
1	Blood	1				
2	Blood	1				
3	Blood	1				
4	Blood	1				
5	Blood	1				
6	Blood	1				

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

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

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



Look at 2021 dataset filtered for IgG

```
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)")
```

2021 dataset IgG PT

Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)

