Pertussis - Mini Project

#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.

```
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,
```

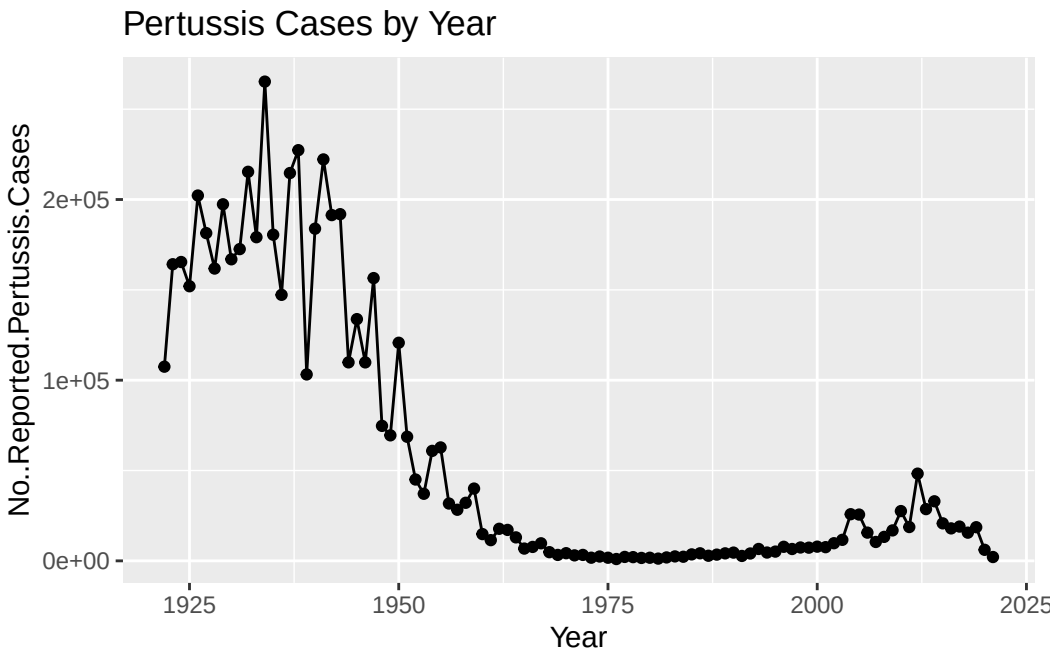
```
                                    13005,6799,7717,9718,4810,3285,4249,
                                    3036,3287,1759,2402,1738,1010,2177,2063,
                                    1623,1730,1248,1895,2463,2276,3589,
                                    4195,2823,3450,4157,4570,2719,4083,6586,
                                    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)
)

cdc
```

```
    Year No..Reported.Pertussis.Cases
1   1922                       107473
2   1923                       164191
3   1924                       165418
4   1925                       152003
5   1926                       202210
6   1927                       181411
7   1928                       161799
8   1929                       197371
9   1930                       166914
10  1931                       172559
11  1932                       215343
12  1933                       179135
13  1934                       265269
14  1935                       180518
15  1936                       147237
16  1937                       214652
17  1938                       227319
18  1939                       103188
19  1940                       183866
20  1941                       222202
21  1942                       191383
22  1943                       191890
23  1944                       109873
24  1945                       133792
25  1946                       109860
26  1947                       156517
27  1948                        74715
28  1949                        69479
```

29	1950	120718
30	1951	68687
31	1952	45030
32	1953	37129
33	1954	60886
34	1955	62786
35	1956	31732
36	1957	28295
37	1958	32148
38	1959	40005
39	1960	14809
40	1961	11468
41	1962	17749
42	1963	17135
43	1964	13005
44	1965	6799
45	1966	7717
46	1967	9718
47	1968	4810
48	1969	3285
49	1970	4249
50	1971	3036
51	1972	3287
52	1973	1759
53	1974	2402
54	1975	1738
55	1976	1010
56	1977	2177
57	1978	2063
58	1979	1623
59	1980	1730
60	1981	1248
61	1982	1895
62	1983	2463
63	1984	2276
64	1985	3589
65	1986	4195
66	1987	2823
67	1988	3450
68	1989	4157
69	1990	4570
70	1991	2719
71	1992	4083

72	1993	6586
73	1994	4617
74	1995	5137
75	1996	7796
76	1997	6564
77	1998	7405
78	1999	7298
79	2000	7867
80	2001	7580
81	2002	9771
82	2003	11647
83	2004	25827
84	2005	25616
85	2006	15632
86	2007	10454
87	2008	13278
88	2009	16858
89	2010	27550
90	2011	18719
91	2012	48277
92	2013	28639
93	2014	32971
94	2015	20762
95	2016	17972
96	2017	18975
97	2018	15609
98	2019	18617
99	2020	6124
100	2021	2116

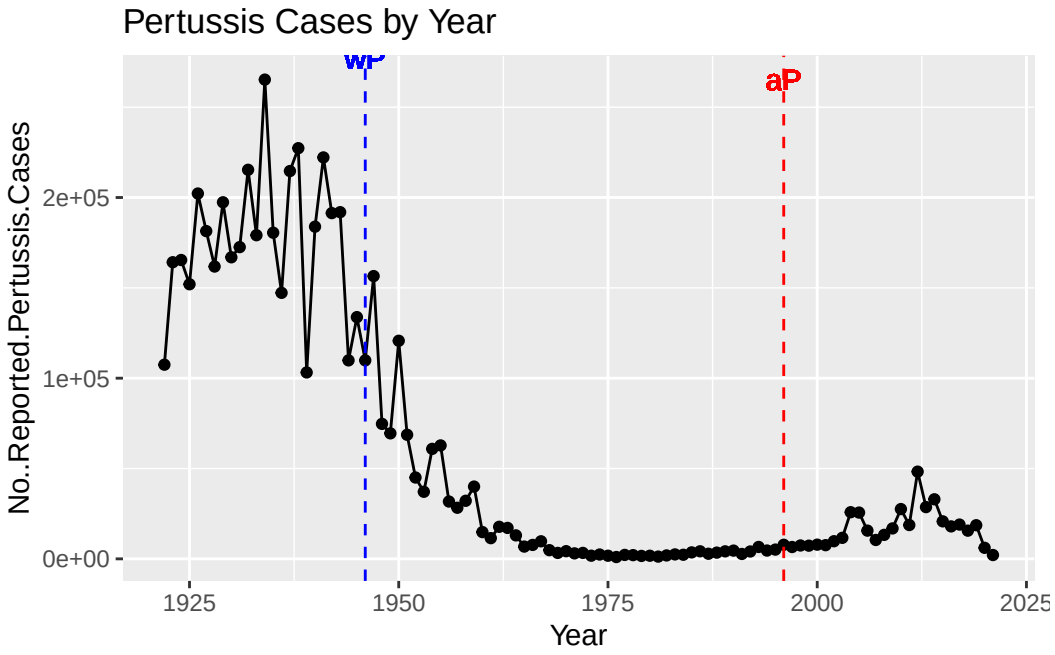
```
library(ggplot2)

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



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 = No..Reported.Pertussis.Cases
) +
  geom_point() +
  geom_line() +
  labs(title = "Pertussis Cases by Year") +
  geom_vline(xintercept = 1946, color = "blue", linetype = "dashed") +
  geom_vline(xintercept = 1996, color = "red", linetype = "dashed") +
  geom_text(aes(x = 1946, label = "wP"), y = max(cdc$No..Reported.Pertussis.Cases), vjust
  geom_text(aes(x = 1996, label = "aP"), y = max(cdc$No..Reported.Pertussis.Cases), color
```



I notice a large decrease after the wP vaccine and a surprising rise after the aP vaccine.

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

There is a rise in the amount of cases when it was almost at 0 in 1975 and onwards, I believe this might be due to vaccine hesitancy or evolution of the bacteria, becoming resistant to the vaccines.

3. Exploring CMI-PB data

```
# 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
```

Q.4 Q4. How many aP and wP infancy vaccinated subjects are in the dataset?

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

```
aP wP
60 58
```

60 aP and 58 wP

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

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

```
Female   Male
    79     39
```

There are 79 females and 39 males

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

```
table(subject$biological_sex, subject$race)
```

```
        American Indian/Alaska Native Asian Black or African American
  Female                             0    21                        2
  Male                               1    11                        0

        More Than One Race Native Hawaiian or Other Pacific Islander
```

```
  Female                   9                                        1
  Male                     2                                        1

         Unknown or Not Reported White
  Female                      11    35
  Male                         4    20
```

1 Native male 21 Asian females and 11 Asian males 2 Black females 9 Mixed females and 2 Mixed Males 1 Pacific Islander male and female 35 white females and 20 white males 11 Unknown females and 4 Unknown males

```
library(lubridate)
```

```
Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

```
#What is today's date?
today()
```

```
[1] "2023-12-11"
```

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

```
Time difference of 8745 days
```

```
#What is this in years?
time_length( today() - ymd("2000-01-01"), "years")
```

```
[1] 23.94251
```

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 todays date to calculate age in days
subject$age <- today() - ymd(subject$year_of_birth)

subject$age
```

```
Time differences in days
  [1] 13858 20433 14954 13128 12032 13128 15684 14223 10206 15319 13858 15319
 [13]  9840 11301 12762 13493 16050  9840 10936 15684 14954 14223 12032 11667
 [25] 13128 14954  9840 15319  9840 13128 12762  9840 12397 14954 12032  9840
 [37]  9475  9840 14223 10936 14223  9840  9475  9475  9840  9475 10206  9475
 [49]  9840  9840  9840  9475  9475  9840  9840  9840 10206  9840  9840  9840
 [61] 13493 11301 10571 11301 12397 17511 18972 18972 12397  9475  9475 12032
 [73] 10571 10571  9475  9475 13128 11301 13493 11667 11301  9475  9110  9840
 [85]  8745  9475  8745  8745  9840  9110  9475  8745 10206  9110  9475  8745
 [97] 13858 11301  9110  8379  7649  7649 10936 12762 10936 10206  9475 10571
[109] 12762  9840 10206 10206 10206 12397  8014  8745 10936  9475
```

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

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     21      26      26      26      27      30
```

```
# wP
wp <- subject %>% filter(infancy_vac == "wP")
round( summary( time_length( wp$age, "years" ) ) )
```

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   28      31      35      36      39      56
```

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

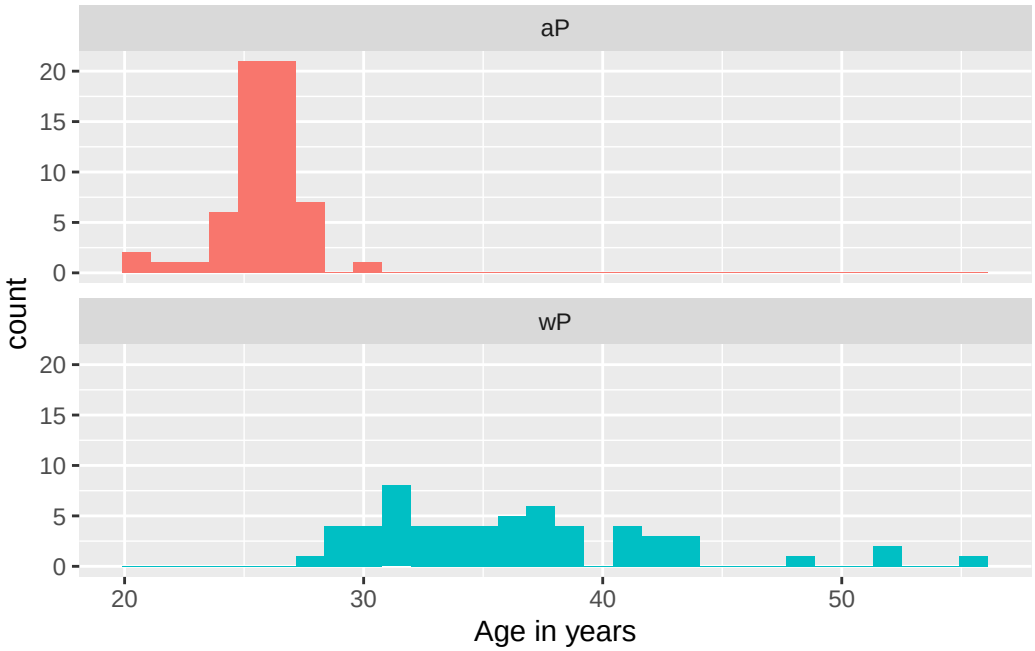
```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

```
[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`.
```



There is a signifcant difference based on the two graphs.

##Joining multiple tables

```
# Assuming you have dplyr loaded
library(dplyr)
library(jsonlite)

# Complete the API URLs...
subject <- read_json("https://www.cmi-pb.org/api/subject", simplifyVector = TRUE)
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] 939  13
```

```
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
```

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] 41810    20
```

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

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

```
 IgE  IgG IgG1 IgG2 IgG3 IgG4
6698 3240 7968 7968 7968 7968
```

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
       31520         8085         2205
```

#4. Examine IgG Ab titer levels

```
igg <- abdata %>% filter(isotype == "IgG")
head(igg)
```

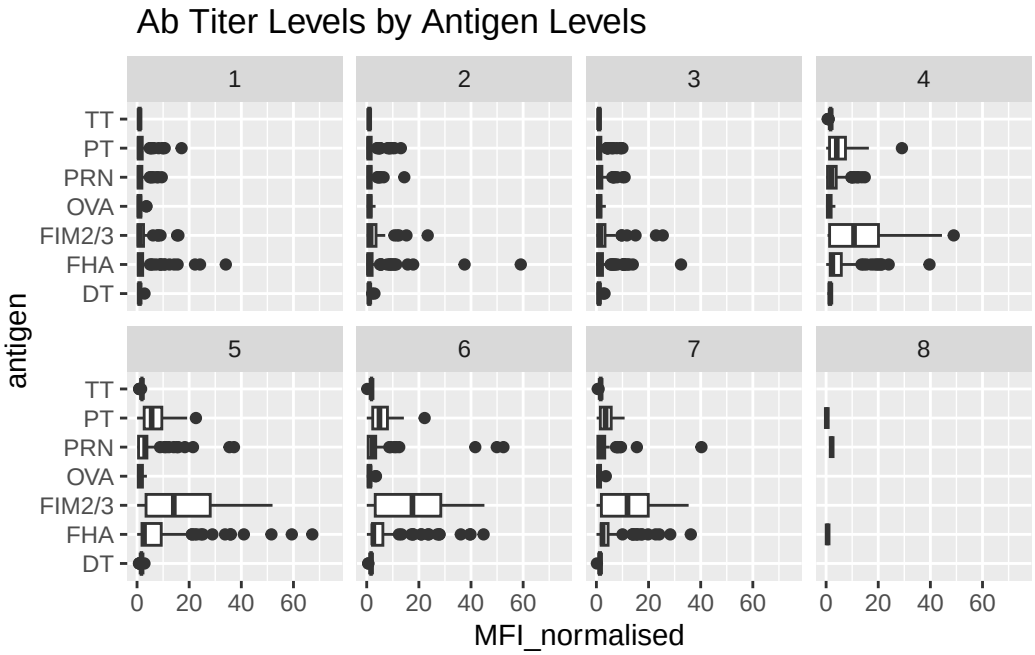
```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1     IgG                TRUE      PT   68.56614       3.736992
2           1     IgG                TRUE     PRN  332.12718       2.602350
3           1     IgG                TRUE     FHA 1887.12263      34.050956
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 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
```

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

```
library(ggplot2)

ggplot(igg) +
  aes(x = MFI_normalised, y = antigen) +
  geom_boxplot() +
  xlim(0, 75) +
  facet_wrap(vars(visit), nrow = 2) +
  labs(title = " Ab Titer Levels by Antigen Levels")
```

```
Warning: Removed 5 rows containing non-finite values (`stat_boxplot()`).
```


Ab Titer Levels by Antigen Levels
1
2
3
4
5
6
7
8
antigen
TT
PT
PRN
OVA
FIM2/3
FHA
DT
0
20
40
60
MFI_normalised