Lab 19: Pertussis Mini Project

Forrest Wang A15911047

#1 Investigating pertussis cases by year

```
library("datapasta")
```

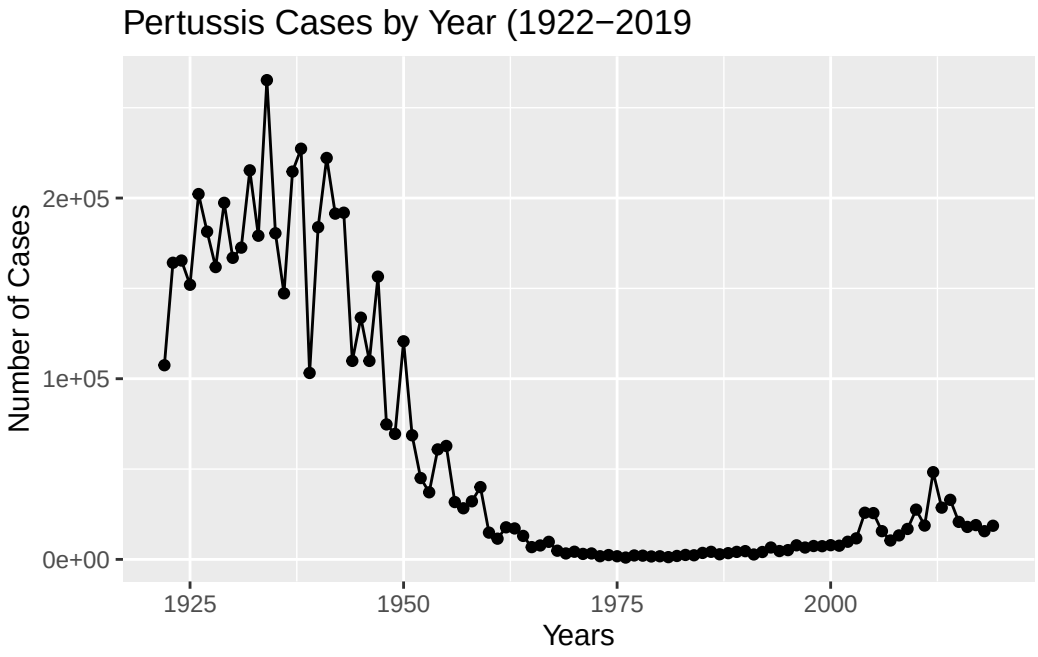
```
library(ggplot2)
```

```
cdc <- data.frame(
                  V1 = 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),
                  V2 = c(107463L,164191L,165418L,
                         152003L,202210L,181411L,161799L,197371L,166914L,
                         172559L,215343L,179135L,265269L,180518L,147237L,
                         214652L,227319L,103188L,183866L,222202L,191383L,191890L,
                         109873L,133792L,109860L,156517L,74715L,69479L,
                         120718L,68687L,45030L,37129L,60886L,62786L,31732L,
                         28295L,32148L,40005L,14809L,11468L,17749L,17135L,
                         13005L,6799L,7717L,9718L,4810L,3285L,4249L,3036L,
                         3287L,1759L,2402L,1738L,1010L,2177L,2063L,1623L,
                         1730L,1248L,1895L,2463L,2276L,3589L,4195L,2823L,
```

```
                    3450L,4157L,4570L,2719L,4083L,6586L,4617L,5137L,
                    7796L,6564L,7405L,7298L,7867L,7580L,9771L,11647L,
                    25827L,25616L,15632L,10454L,13278L,16858L,27550L,
                    18719L,48277L,28639L,32971L,20762L,17972L,18975L,
                    15609L,18617L)
        )
```

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(x = V1, y = V2) +
  geom_point() +
  geom_line() +
  ggtitle("Pertussis Cases by Year (1922-2019") +
  xlab("Years") +
  ylab("Number of Cases")
```



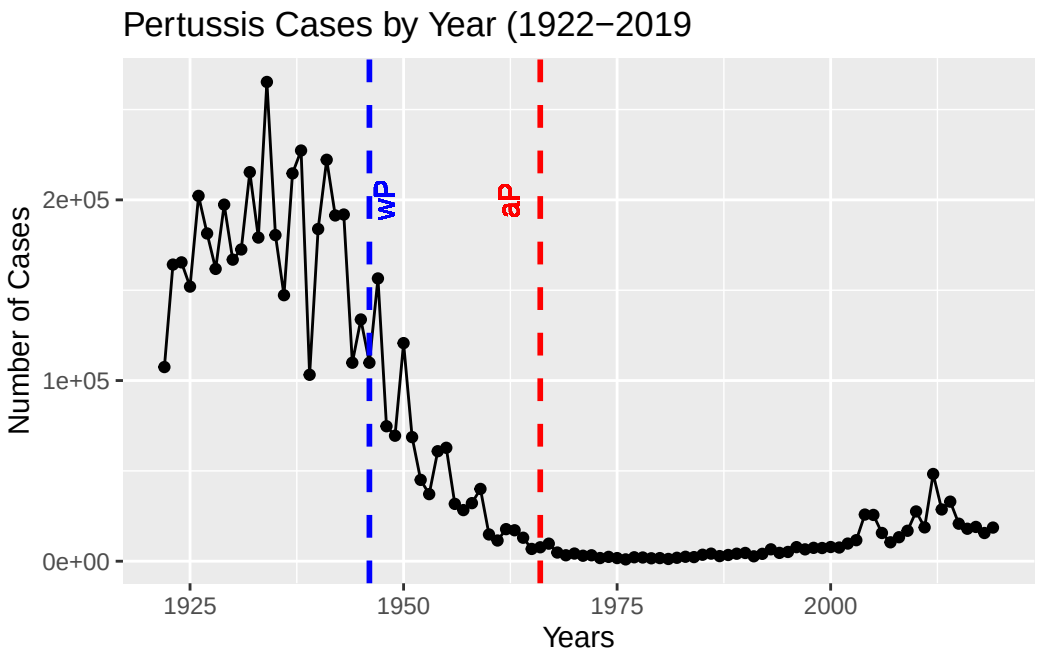
#2. A tale of two vaccines (wP & aP)

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 = V1, y = V2) +
  geom_point() +
  geom_line() +
  ggtitle("Pertussis Cases by Year (1922-2019") +
  xlab("Years") +
  ylab("Number of Cases") +
  geom_vline(xintercept=1946, linetype="dashed",
               color = "blue", size=1)+
  geom_vline(xintercept=1966, linetype="dashed",
               color = "red", size=1)+
  geom_text(aes(x=1946, label="wP", y=2e+05), colour="blue", angle=90, vjust = 1.2, text=e
  geom_text(aes(x=1966, label="aP", y=2e+05), colour="red", angle=90, vjust = -1, text=ele
```

```
Warning: Ignoring unknown parameters: text
Ignoring unknown parameters: text
```



Q3. Describe what happened after the introduction of the aP vaccine? Do you have a possible explanation for the observed trend? A3. Pertussis cases per year were at a steady decline

following wP introducction. After 1966 with aP introduction, cases were very low, below 5,000 per year up until 1993. A possible explanation for this period of low cases are due to the fact that aP used PURE ANTIGENS and with LESS SIDE EFFECTS, encouraging more vaccine use.

HOWEVER, it's RESURGANCE may be caused by bacterial immunity overtime to the aP vaccine and vaccine hessitancy

#3. Exploring CMI-PB 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 may aP and wP infancy vaccinated subjects are in the dataset?

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

```
aP wP
47 49
```

A4. There are a total of 96 infancy vaccinated subjects in the dataset, 47 of which received wP, and 49 aP

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

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

```
Female   Male
    66     30
```

A5: There are 30 male subjects and 66 female subjects

Q6. What is the breakdown of race and biological sex (e.g. number of Asian females, White males etc…)? A6:

```
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
```

#Side-Note: Working with dates

```
library(lubridate)
```

```
Loading required package: timechange
```

```

Attaching package: 'lubridate'
```

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

    date, intersect, setdiff, union
```

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

```
[1] 22.9295
```

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? A7: (i) & (ii)

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

```
Time differences in days
 [1] 13488 20063 14584 12758 11662 12758 15314 13853  9836 14949 13488 14949
[13]  9470 10931 12392 13123 15680  9470 10566 13123 10931 10201 10931 12027
[25] 17141 18602 18602 12027  9105  9105 11662 10201 10201  9105  9105 12758
[37] 10931 13123 11297 10931  9105  8740  9470  8375  9105  8375  8375  9470
[49]  8740  9105  8375  9836  8740  9105  8375 15314 14584 13853 11662 11297
[61] 12758 14584  9470 14949  9470 12758 12392  9470 12027 14584 11662  9470
[73]  9105  9470 13853 10566 13853  9470  9105  9105  9470  9105  9836  9105
[85]  9470  9470  9470  9105  9105  9470  9470  9470  9836  9470  9470  9470
```

```
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.
     23      25      26      25      26      27
```

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

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

(iii) The average age of aP is 25 while wP is 36

Q8. Determine the age of all individuals at time of boost? A9:

```
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
```

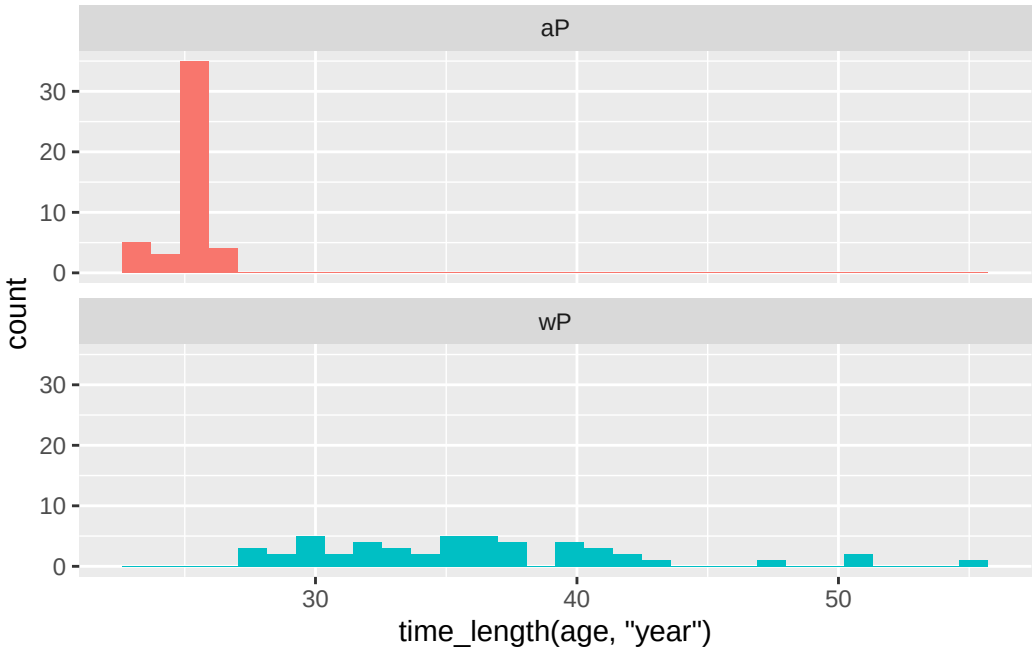
```
int/365
```

```
Time differences in days
 [1] 30.71781 51.10959 33.79726 28.67945 25.67671 28.79452 35.87397 34.17260
 [9] 20.57808 34.58630 30.67945 34.58630 19.57534 23.63562 27.63836 29.58356
[17] 36.72329 19.67123 22.75068 32.28767 25.91781 23.91781 25.91781 28.93973
[25] 42.95068 47.10685 47.10685 29.09315 21.08767 21.08767 28.16986 24.16712
[33] 24.16712 21.16438 21.16438 31.23014 26.22466 32.23014 27.22740 26.22466
[41] 21.22192 20.27945 22.27945 19.33699 21.33699 19.33699 19.33699 22.43288
[49] 20.43288 21.43288 19.49041 23.49315 20.49041 21.49041 19.49041 35.68219
[57] 33.68219 31.67945 25.75342 24.71781 28.72055 33.75890 19.74795 34.75890
[65] 19.74795 28.75616 27.75342 19.82466 26.79178 33.83562 25.79178 19.82466
[73] 18.86301 19.82466 31.83288 22.82740 31.87123 19.86301 18.86301 18.86301
[81] 19.92055 18.86301 20.92329 19.05753 20.05753 19.92055 19.92055 19.01918
[89] 19.01918 20.05753 20.05753 20.09315 21.09589 20.09315 20.09315 20.09315
```

Q9. With the help of a faceted boxplot (see below), do you think these two groups are significantly different? A9. According to the boxplot below, and because p-value is less than 0.5, there is a significance difference between the 2 groups

```
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
```

#Joining multiple tables

```
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)
```

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 <- full_join(specimen, subject)
```

```
Joining, by = "subject_id"
```

```
dim(meta)
```

```
[1] 729  14
```

```
head(meta)
```

```
  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 infancy_vac biological_sex
1                             0         Blood     1          wP         Female
2                           736         Blood    10          wP         Female
3                             1         Blood     2          wP         Female
4                             3         Blood     3          wP         Female
5                             7         Blood     4          wP         Female
6                            14         Blood     5          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 13488 days
2 13488 days
3 13488 days
4 13488 days
5 13488 days
6 13488 days
```

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?

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

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

A11. For each isotype (IgE IgG IgG1 IgG2 IgG3 IgG4), their respetive entries in abdata are: 6698, 1413, 6141, 6141, 6141, & 6141

Q12. What do you notice about the number of visit 8 specimens compared to other visits?
A12: Compared to the other visits, the number of visit 8 speciments are drastically lower, in the 2 digits as opposed to 4 digits

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

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

#4. Examine IgG1 Ab titer levels

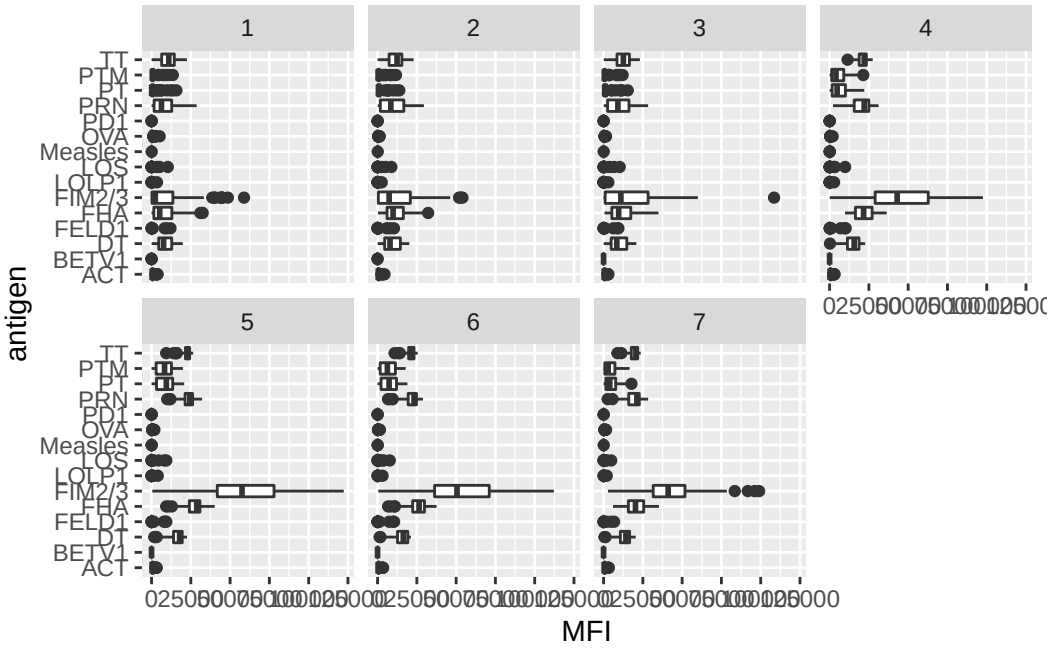
```
ig1 <- abdata %>% filter(isotype == "IgG1", visit!=8)
head(ig1)
```

```
  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 13488 days
2 13488 days
3 13488 days
4 13488 days
5 13488 days
6 13488 days
```

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? A14. FIM2/3 shows differences in level of IgG1 antibody titers overtime. Others do not show any difference as their MFI is much lower.