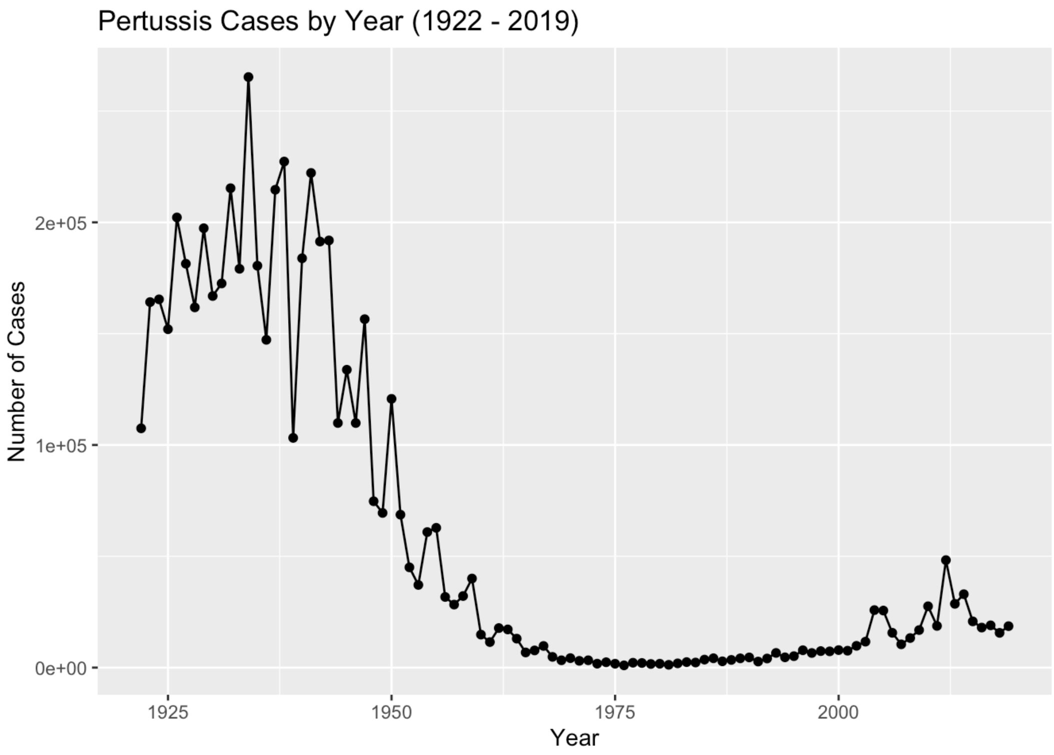
11_30_22_lab

AUTHOR
Andy Tong

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.

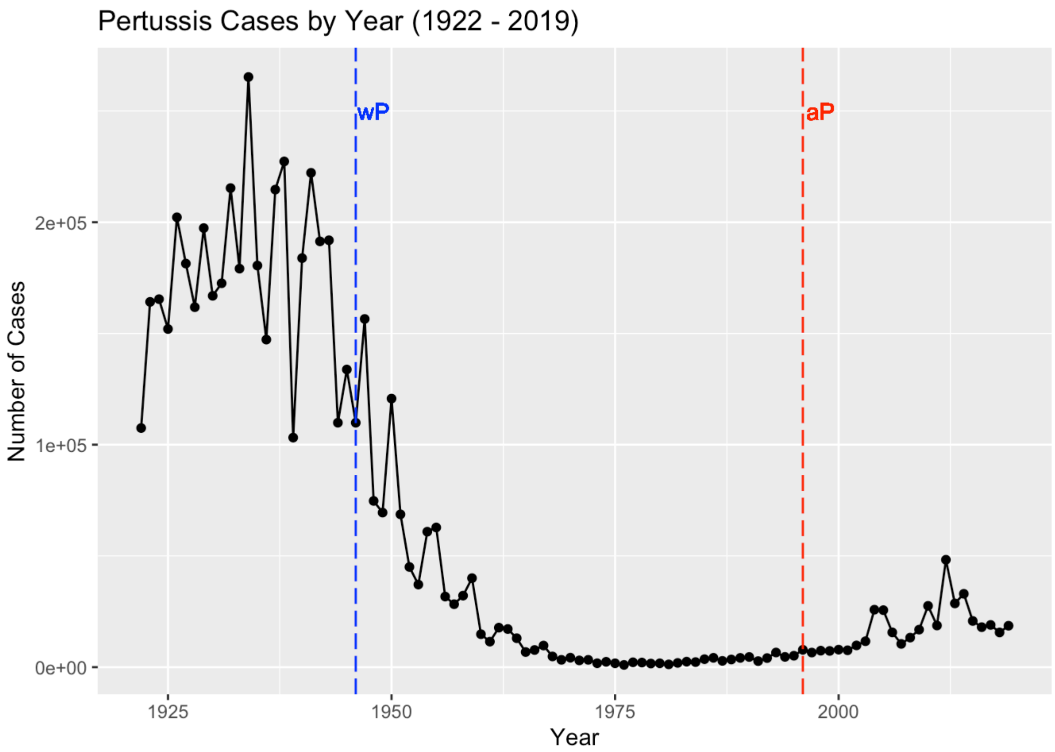
```
pertussis_cases_by_year <- data.frame(
                          Year = c(1922L,
                                   1923L,1924L,1925L,1926L,1927
                                   1929L,1930L,1931L,1932L,1933
                                   1936L,1937L,1938L,1939L,1940
                                   1942L,1943L,1944L,1945L,1946
                                   1949L,1950L,1951L,1952L,1953
                                   1955L,1956L,1957L,1958L,1959
                                   1961L,1962L,1963L,1964L,1965
                                   1968L,1969L,1970L,1971L,1972
                                   1974L,1975L,1976L,1977L,1978
                                   1981L,1982L,1983L,1984L,1985
                                   1987L,1988L,1989L,1990L,1991
                                   1994L,1995L,1996L,1997L,1998
                                   2000L,2001L,2002L,2003L,2004
                                   2006L,2007L,2008L,2009L,2010
                                   2013L,2014L,2015L,2016L,2017
                                   2019L),
  No..Reported.Pertussis.Cases = c(107473,
                                   164191,165418,152003,202210,
                                   161799,197371,166914,172559,
                                   265269,180518,147237,214652,
                                   183866,222202,191383,191890,
                                   133792,109860,156517,74715,6
                                   68687,45030,37129,60886,6278
                                   32148,40005,14809,11468,1774
                                   13005,6799,7717,9718,4810,32
                                   3036,3287,1759,2402,1738,101
                                   1623,1730,1248,1895,2463,227
                                   4195,2823,3450,4157,4570,271
                                   4617,5137,7796,6564,7405,729
                                   7580,9771,11647,25827,25616,
                                   13278,16858,27550,18719,4827
                                   32971,20762,17972,18975,1560
)
```

```
library(ggplot2)
ggplot(pertussis_cases_by_year) +
  aes(Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +
  labs(x = "Year", y="Number of Cases", title = "Pertussis Case
```



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(pertussis_cases_by_year) +
  aes(Year, No..Reported.Pertussis.Cases) +
  geom_point() +
  geom_line() +
  labs(x = "Year", y="Number of Cases", title = "Pertussis Case
  geom_vline(xintercept = c(1946, 1996),  colour=c("blue", "rec
  geom_text(aes(x = 1946+2, y = 250000, label = "wP"), colour =
  geom_text(aes(x = 1996+2, y = 250000, label = "aP"), colour =
```



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 the aP vaccine there was a slight uptick in cases of Pertussis. It did take a while though, this gives credence to maybe it was due to other factors such as testing becoming more sensitive and reporting more cases. Another possibility is the vaccine is slightly weaker but gives fewer side effects in a cost benefit analysis concluding that it is better. Another explanation is if the vaccine isn’t updating often enough, the bacterial evolution could outpace vaccine development, creating new strains that could get around vaccination. A final explanation is that with time, people begin fearing the whooping cough less as they have never seen anyone get it. Because of this, they don’t vaccinate their children for it as it doesn’t seem like a problem. This could lead to a rise in cases. Between the fall in cases to almost 0 to the uptick, there are about 30 years, enough for one generation to grow up without fear or respect of the disease.

```
library(jsonlite)
subject <- read_json("https://www.cmi-pb.org/api/subject", simp
```

```
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?

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

```
aP wP
47 49
```

47 aP infancy vaccinated subjects and 49 wP infancy vaccinated subjects

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

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

```
Female   Male
    66     30
```

66 females and 30 males are in the dataset

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                                         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
```

Here is the breakdown

```r
library(lubridate)
```

```
Loading required package: timechange


Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

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?

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

```r
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
```

```r
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 <- 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 
```

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

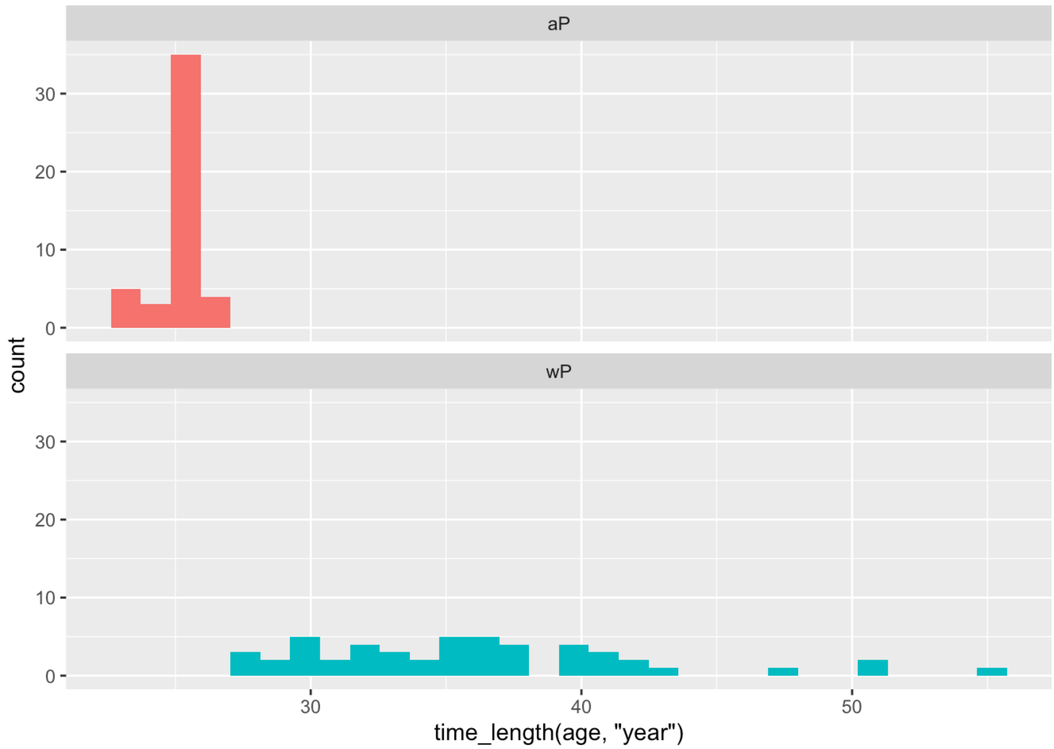
```
time_diff_recieved_boost <- ymd(subject$date_of_boost) - ymd(su
age_at_boost <- time_length(time_diff_recieved_boost, "year")
head(age_at_boost)
```

```
[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)
```

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



Yes they look significantly different. In the past people seem to get the vaccine whenever they could, but not at a young age. This leads to a distribution ranging from 25ish to almost 60. With the new vaccine everyone gets it before 30.

```
# Complete the API URLs...
specimen <- read_json("https://www.cmi-pb.org/api/specimen", si
titer <- read_json("https://www.cmi-pb.org/api/ab_titer", simpl
```

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 13483 days
2 13483 days
3 13483 days
4 13483 days
5 13483 days
6 13483 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 
```

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

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

```

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

There are very few visit 8 specimens compared to all the other visits.

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

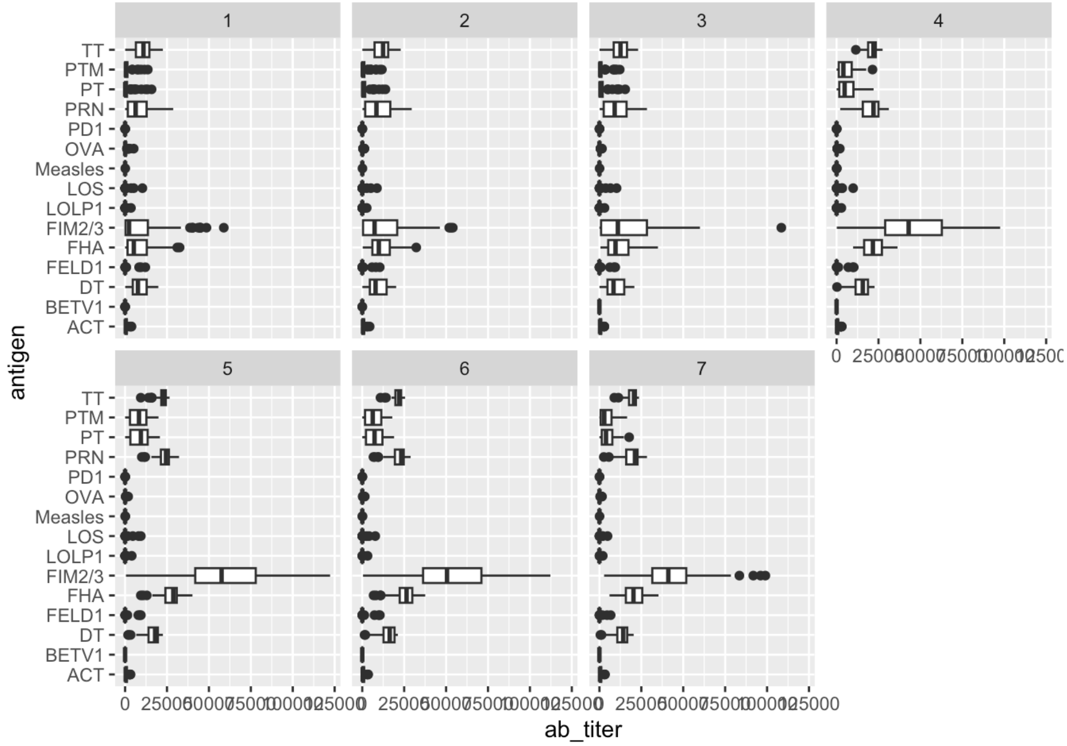
```
  specimen_id isotype is_antigen_specific antigen        MFI
MFI_normalised
1           1    IgG1                TRUE     ACT 274.355068
0.6928058
2           1    IgG1                TRUE     LOS  10.974026
2.1645083
3           1    IgG1                TRUE   FELD1   1.448796
0.8080941
4           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 13483 days
2 13483 days
3 13483 days
4 13483 days
5 13483 days
6 13483 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)+
  labs(x = "ab_titer")
```



Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others?

FIM2/3 seems the most unique in terms of the level of IgG1 antibody titers. I didn’t find anything on the website so I’m not entirely sure what that means. I am somewhat confused.