class 18

Q1)

```
options(repos = c(CRAN = "https://cran.rstudio.com"))
```

```
install.packages("ggplot2")
```

```

The downloaded binary packages are in
    /var/folders/8v/ljcn64zd6bs7yszj9yy3lnc80000gn/T//RtmpAfVhGI/downloaded_packages
```

```
install.packages("datapasta")
```

```

The downloaded binary packages are in
    /var/folders/8v/ljcn64zd6bs7yszj9yy3lnc80000gn/T//RtmpAfVhGI/downloaded_packages
```

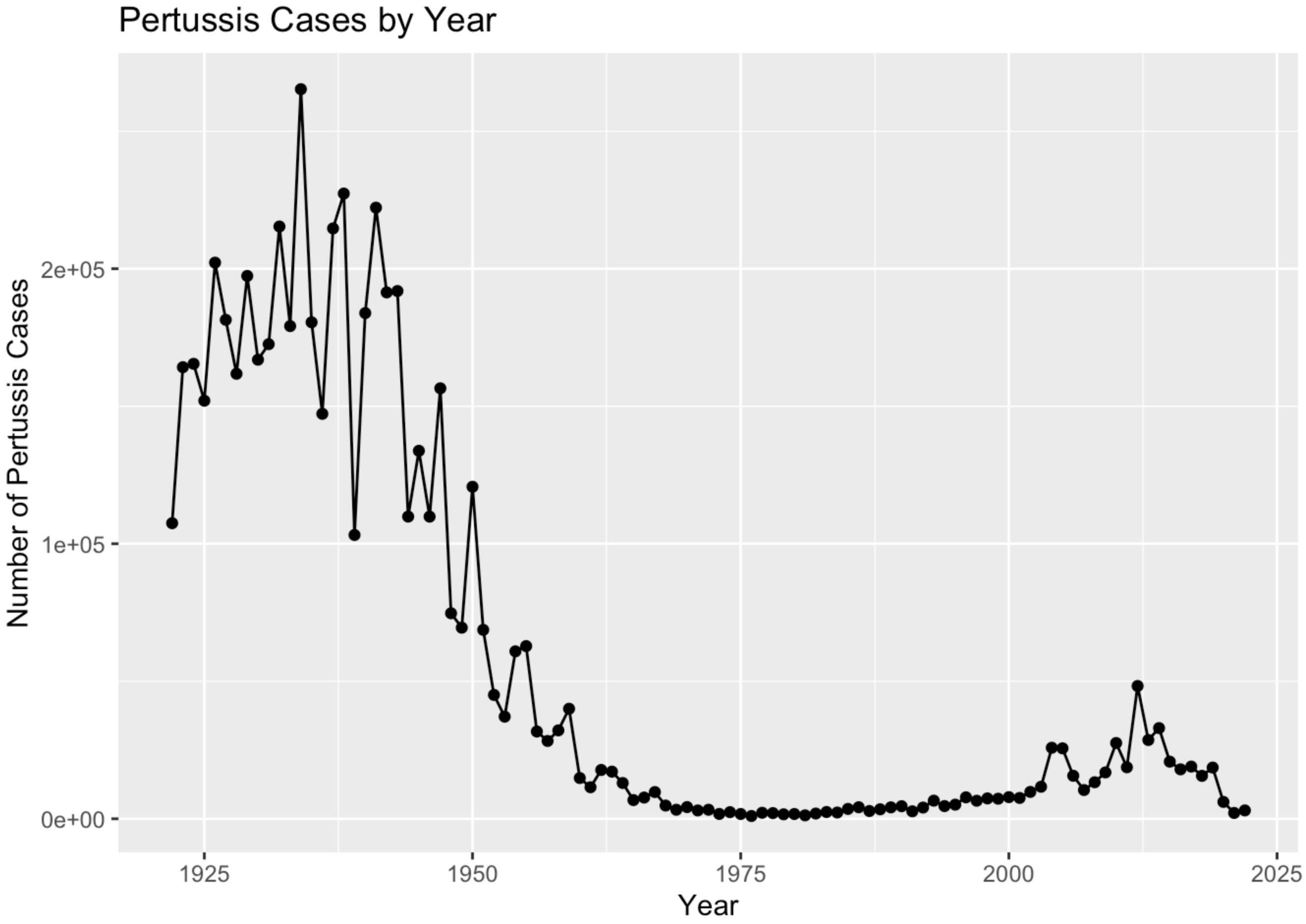
```
library(ggplot2)
library(datapasta)
```

```
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,2022L),
  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,
                                   2116,3044)
)
```

```
head(cdc)
```

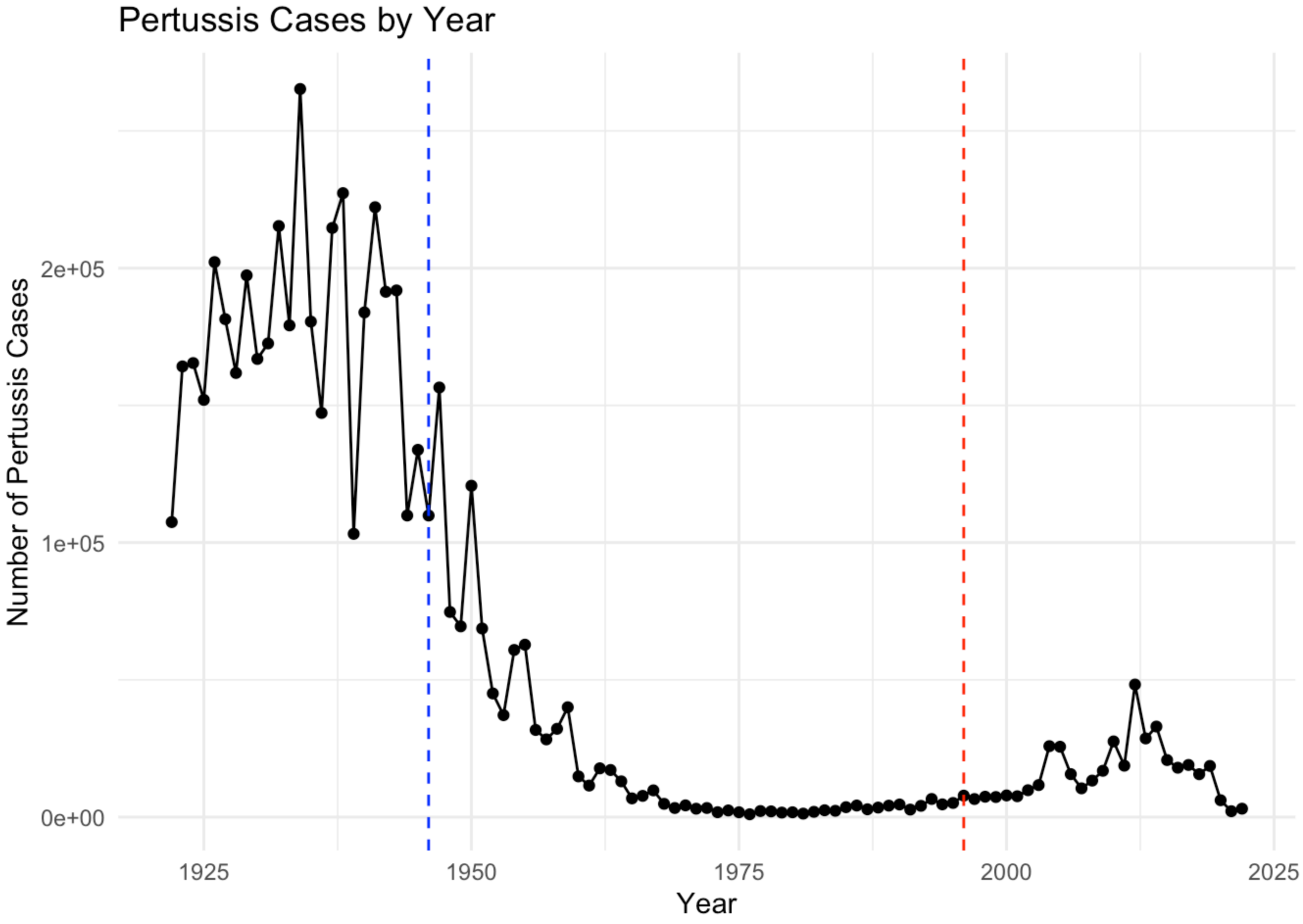
```
  Year No..Reported.Pertussis.Cases
1 1922                       107473
2 1923                       164191
3 1924                       165418
4 1925                       152003
5 1926                       202210
6 1927                       181411
```

```
ggplot(cdc) +
  aes(x = Year, y = No..Reported.Pertussis.Cases) +  # Adjust column names as needed
  geom_point() +  # Scatter plot
  geom_line() +  # Line plot
  labs(title = "Pertussis Cases by Year", x = "Year", y = "Number of Pertussis Cases")
```



Q2)

```
ggplot(cdc) +
  aes(x = Year, y = No..Reported.Pertussis.Cases) +  # Plot pertussis cases by year
  geom_point() +  # Scatter plot
  geom_line() +  # Line plot
  geom_vline(xintercept = 1946, linetype = "dashed", color = "blue") +  # wP vaccine intr
  geom_vline(xintercept = 1996, linetype = "dashed", color = "red") +   # aP vaccine swit
  labs(
    title = "Pertussis Cases by Year",
    x = "Year",
    y = "Number of Pertussis Cases"
  ) +
  theme_minimal()  # Optional: Adds a cleaner theme
```



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

Pertussis cases increase again. In 2012 there were 48,277 cases in the US. This is the largest number of cases reported since 1955, where there were 62,786 cases.

This could be due to more sensitive PCR-based testing, vaccination hesitancy, or bacterial evolution.

```
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 many aP and wP infancy vaccinated subjects are in the dataset?

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

```

aP wP 
87 85 
```

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

```
# Count the number of Male and Female subjects
table(subject$biological_sex)
```

```

Female   Male 
   112     60 
```

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                                         32   12
  Black or African American                      2    3
  More Than One Race                            15    4
  Native Hawaiian or Other Pacific Islander      1    1
  Unknown or Not Reported                       14    7
  White                                         48   32
```

```
library(lubridate)
```

```

Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

```
today()
```

```
[1] "2025-03-08"
```

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

```
Time difference of 9198 days
```

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

```
[1] 25.18275
```

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?

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

```
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. 
     22      26      27      27      28      34 
```

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

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     22      32      34      36      39      57 
```

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

```
# Load necessary library
library(lubridate)

# Convert date_of_boost to Date and year_of_birth to a full date (e.g., January 1st of th
subject$date_of_boost <- ymd(subject$date_of_boost)  # Assuming it's already in 'YYYY-MM-
subject$year_of_birth <- ymd(paste(subject$year_of_birth, "01", "01", sep = "-"))  # Crea
```

```
Warning: All formats failed to parse. No formats found.
```

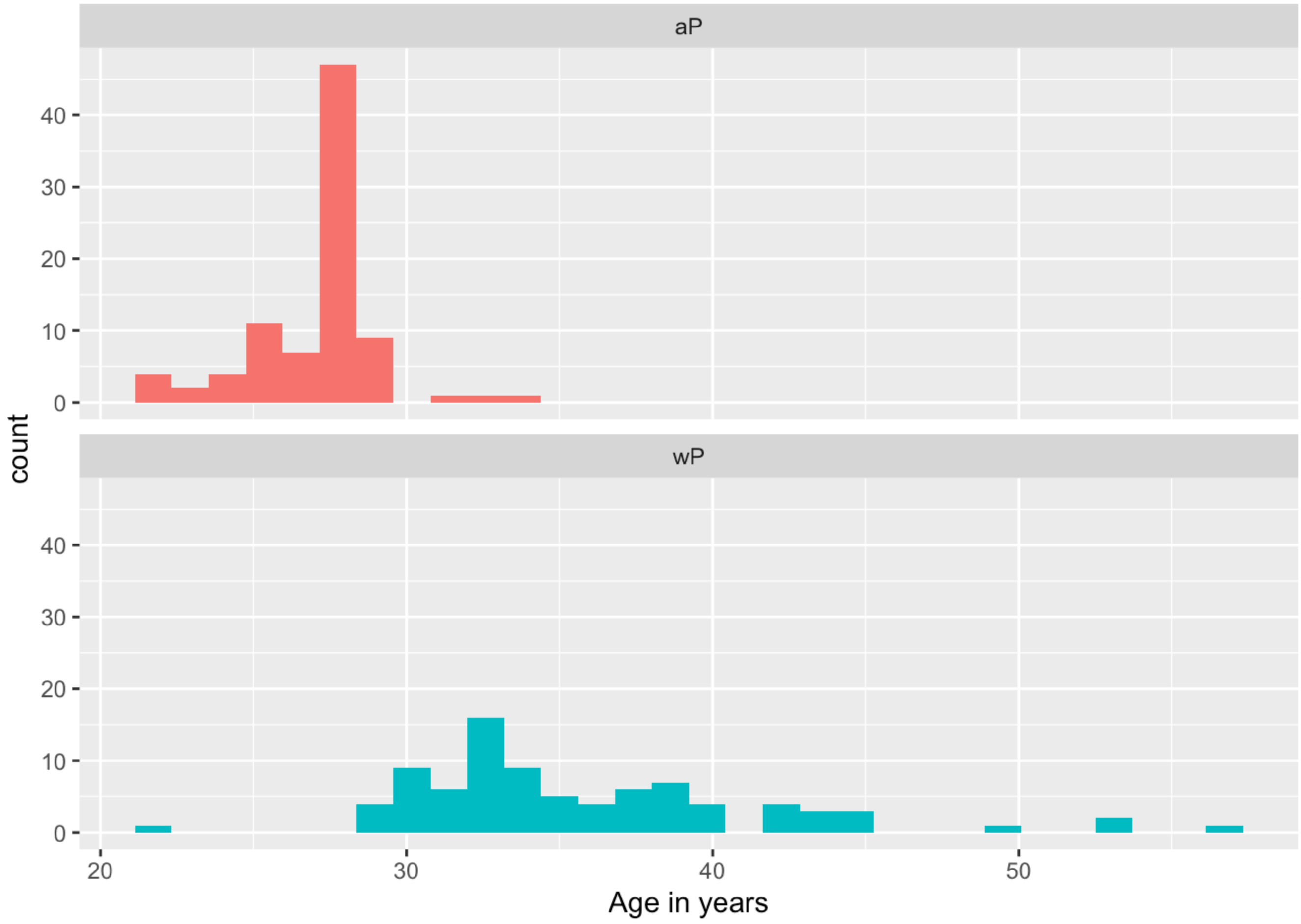
```
# Calculate the difference between boost date and birth date
int <- subject$date_of_boost - subject$year_of_birth

# Convert time difference to years
age_at_boost <- time_length(int, "year")

# Display the result
head(age_at_boost)
```

```
[1] NA NA NA NA NA NA
```

With the help of a faceted boxplot or histogram (below), do you think these two groups are significantly different? Yes they are statistacally 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`.
```



```r
install.packages("dplyr")
```

```
The downloaded binary packages are in
	/var/folders/8v/ljcn64zd6bs7yszj9yy3lnc80000gn/T//RtmpAfVhGI/downloaded_packages
```

```r
install.packages("jsonlite")
```

```
The downloaded binary packages are in
	/var/folders/8v/ljcn64zd6bs7yszj9yy3lnc80000gn/T//RtmpAfVhGI/downloaded_packages
```

```r
library(jsonlite)
specimen <- read_json("https://www.cmi-pb.org/api/v5_1/specimen", simplifyVector = TRUE)
titer <- read_json("https://www.cmi-pb.org/api/v5_1/plasma_ab_titer", simplifyVector = TR
```

9. 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:

```r
library(dplyr)
meta <- inner_join(subject,specimen)
```

```
Joining with `by = join_by(subject_id)`
```

```r
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          <NA>    2016-09-12 2020_dataset 14311 days           1
2          <NA>    2016-09-12 2020_dataset 14311 days           2
3          <NA>    2016-09-12 2020_dataset 14311 days           3
4          <NA>    2016-09-12 2020_dataset 14311 days           4
5          <NA>    2016-09-12 2020_dataset 14311 days           5
6          <NA>    2016-09-12 2020_dataset 14311 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. 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] 61956    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  7265 11993 12000 12000 12000 
```

```
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          <NA>    2016-09-12 2020_dataset
2 Not Hispanic or Latino White          <NA>    2016-09-12 2020_dataset
3 Not Hispanic or Latino White          <NA>    2016-09-12 2020_dataset
4                Unknown White          <NA>    2016-10-10 2020_dataset
5                Unknown White          <NA>    2016-10-10 2020_dataset
```

```
6                Unknown White          <NA>    2016-10-10 2020_dataset
         age actual_day_relative_to_boost planned_day_relative_to_boost
1 14311 days                           -3                             0
2 14311 days                           -3                             0
3 14311 days                           -3                             0
4 15407 days                           -3                             0
5 15407 days                           -3                             0
6 15407 days                           -3                             0
  specimen_type visit
1         Blood     1
2         Blood     1
3         Blood     1
4         Blood     1
5         Blood     1
6         Blood     1
```

Q12. What are the different $dataset values in abdata and what do you notice about the number of rows for the most “recent” dataset?

```
#  the unique values of dataset column
unique(abdata$dataset)
```

```
[1] "2020_dataset" "2021_dataset" "2022_dataset" "2023_dataset"
```

```
#  the number of rows for each dataset
table(abdata$dataset)
```

```

2020_dataset 2021_dataset 2022_dataset 2023_dataset 
       31520         8085         7301        15050 
```

Data collection decreased significantly after 2020.

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

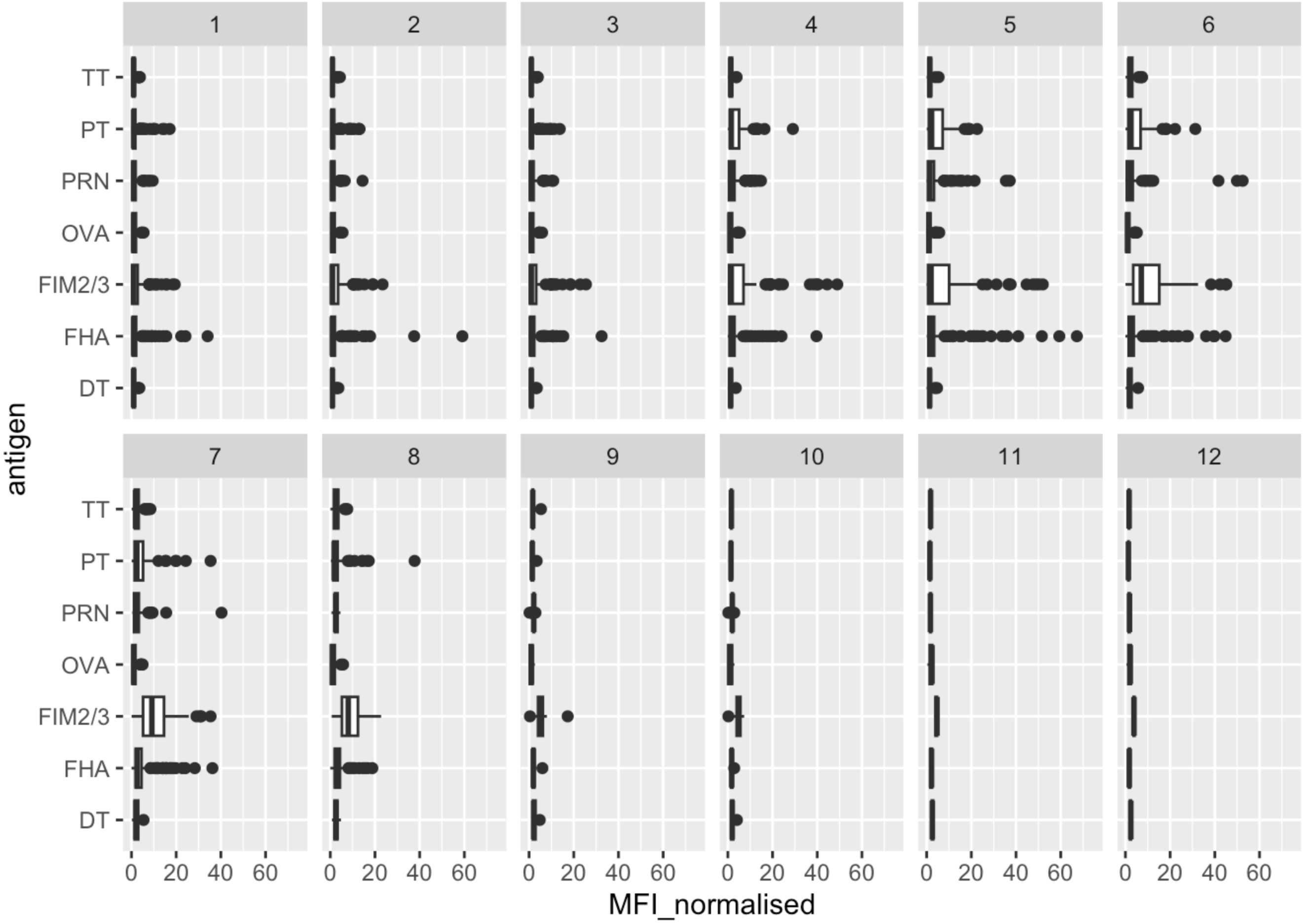
```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1     IgG                TRUE      PT   68.56614       3.736992
2           1     IgG                TRUE     PRN  332.12718       2.602350
3           1     IgG                TRUE     FHA 1887.12263      34.050956
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          <NA>    2016-09-12 2020_dataset
2 Not Hispanic or Latino White          <NA>    2016-09-12 2020_dataset
3 Not Hispanic or Latino White          <NA>    2016-09-12 2020_dataset
4                  Unknown White          <NA>    2016-10-10 2020_dataset
5                  Unknown White          <NA>    2016-10-10 2020_dataset
6                  Unknown White          <NA>    2016-10-10 2020_dataset
         age actual_day_relative_to_boost planned_day_relative_to_boost
1 14311 days                           -3                             0
2 14311 days                           -3                             0
3 14311 days                           -3                             0
4 15407 days                           -3                             0
5 15407 days                           -3                             0
6 15407 days                           -3                             0
  specimen_type visit
1         Blood     1
2         Blood     1
3         Blood     1
4         Blood     1
5         Blood     1
6         Blood     1
```

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

```
ggplot(igg) +
  aes(x = MFI_normalised, y = antigen) +   # Use MFI_normalised for the x-axis
  geom_boxplot() +                          # Create a boxplot
  xlim(0, 75) +                             # Set the x-axis limits from 0 to 75
  facet_wrap(vars(visit), nrow = 2)         # Facet by 'visit' in 2 rows
```

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



Q14. What antigens show differences in IgG antibody titer levels t them over time? Why these and not others?

PRN, FHA, FIM 2/3, PT. Those are actually a part of the vaccine. The others are control.