Class 19 mini project Pertussis

Trinity Lee A16639698

Pertussis is a sever lung infection also known as whooping cough.

We will begin by investigating the number of pertussis cases per year in the US.

This data is available on the CDC website here

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.

```
#/ echo=FALSE

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),
       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)
    )
```

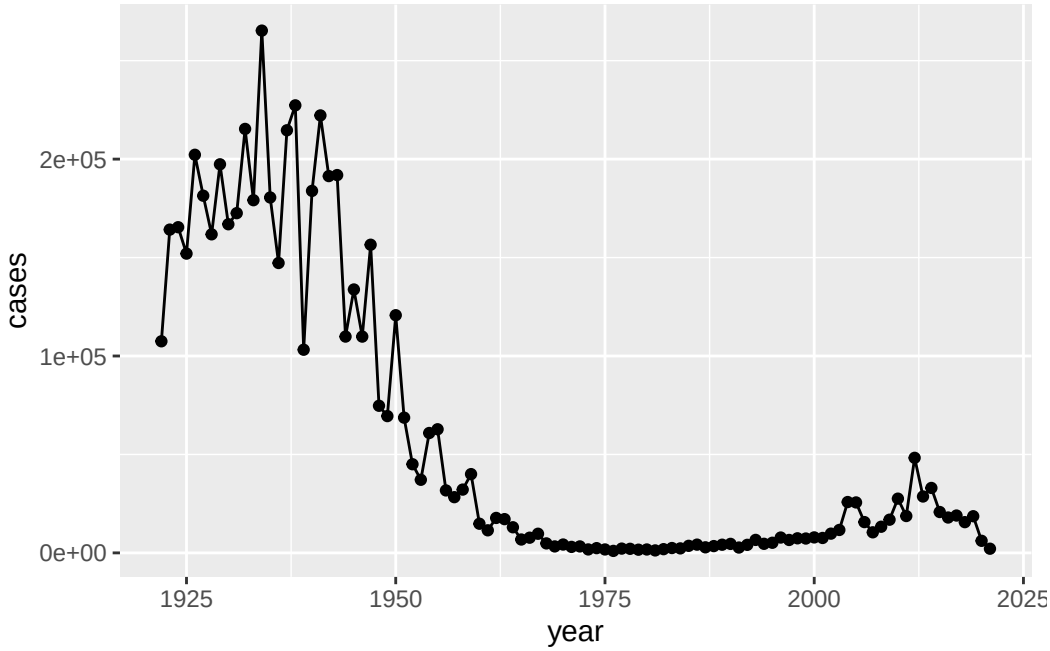
Lets have a look at this data.frame

```
head(cdc)
```

```
  year  cases
1 1922 107473
2 1923 164191
3 1924 165418
4 1925 152003
5 1926 202210
6 1927 181411
```

I want to make a nice plot of cases per year

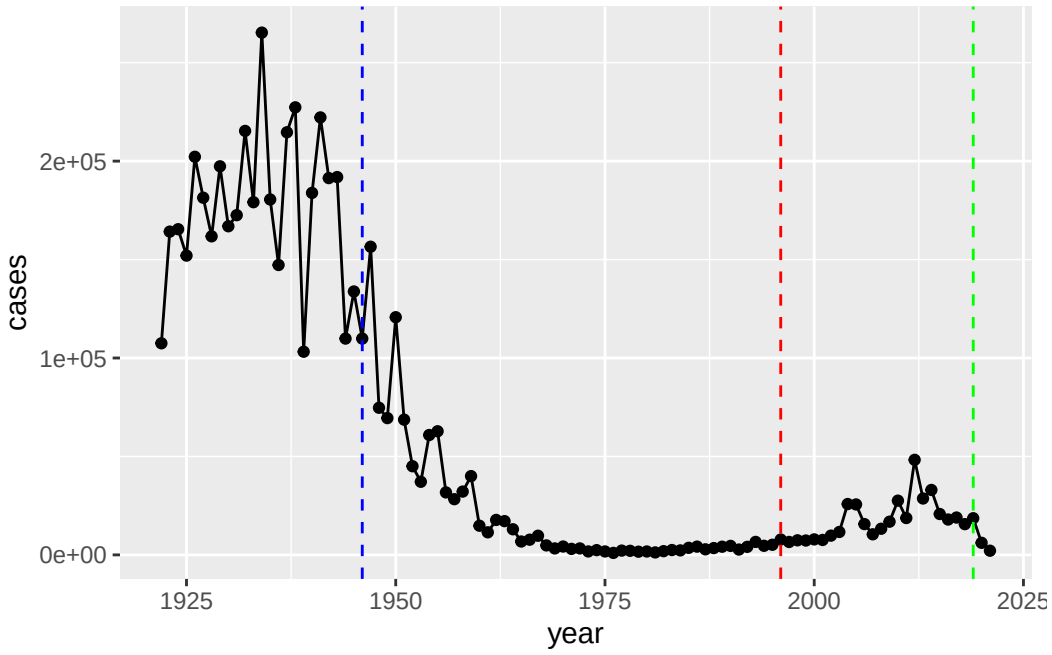
```
library(ggplot2)
ggplot(cdc)+aes(year,cases)+geom_point()+geom_line()
```



#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?

```
library(ggplot2)
ggplot(cdc)+aes(year,cases)+geom_point()+geom_line()+geom_vline(xintercept=1946, linetype=
```



There is more immunity seen with the aP vaccine than the wP vaccine seen by the significantly lower number of cases in 1996 compared to 1946.

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 we see an increase of cases emerging in later years most likely due to bacterial evolution of the reluctancy to vaccinate children with emerging false information.

##3. Exploring CMI-PB data

Why is this vaccine-preventable disease on the upswing? To answer this question we need to investigate the mechanisms underlying waning protection against pertussis. This requires evaluation of pertussis-specific immune responses over time in wP and aP vaccinated individuals.

This is the goals of the CMI-PB project (https://www.cmi-pb.org/)

The CMI-PB project makes its data available via "API-endpoint" that return JSON format. We will use the `jsonlite` package to access this data. The main function in this package is `read_json()`

```
# Allows us to read, write and process JSON data
library(jsonlite)
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)
```

Have peek at new objects

```
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
```

```
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
```

```
  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
```

Q4. How may 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
```

#Working with dates Two of the columns of `subject` contain dates in the Year-Month-Day format. Recall from our last mini-project that dates and times can be annoying to work with at the best of times. However, in R we have the excellent lubridate package, which can make life allot easier.

```
library(lubridate)
```

```
Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

```
today()
```

```
[1] "2023-12-05"
```

```
time_length(today()-mdy("05-31-2002"),"years")
```

```
[1] 21.51403
```

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)
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 <- 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
```

There seems to be a signifcant difference between aP and wP 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")
age_at_boost
```

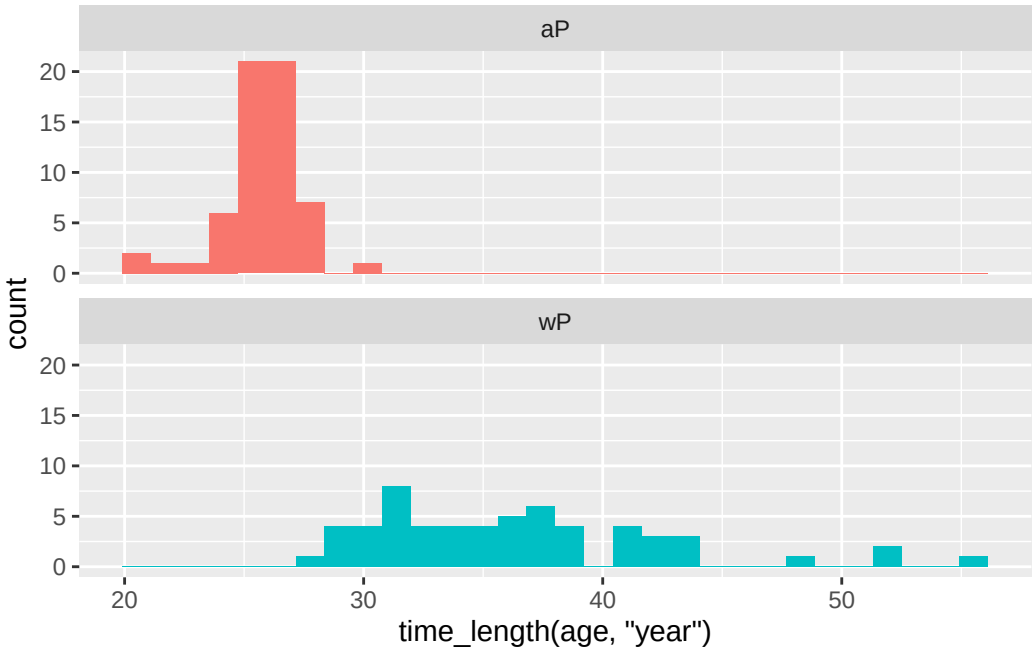
```
  [1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481 35.84942 34.14921
  [9] 20.56400 34.56263 30.65845 34.56263 19.56194 23.61944 27.61944 29.56331
 [17] 36.69815 19.65777 22.73511 35.65777 33.65914 31.65777 25.73580 24.70089
 [25] 28.70089 33.73580 19.73443 34.73511 19.73443 28.73648 27.73443 19.81109
 [33] 26.77344 33.81246 25.77413 19.81109 18.85010 19.81109 31.81109 22.81177
 [41] 31.84942 19.84942 18.85010 18.85010 19.90691 18.85010 20.90897 19.04449
```

```
 [49] 20.04381 19.90691 19.90691 19.00616 19.00616 20.04381 20.04381 20.07940
 [57] 21.08145 20.07940 20.07940 20.07940 32.26557 25.90007 23.90144 25.90007
 [65] 28.91992 42.92129 47.07461 47.07461 29.07324 21.07324 21.07324 28.15058
 [73] 24.15058 24.15058 21.14990 21.14990 31.20876 26.20671 32.20808 27.20876
 [81] 26.20671 21.20739 20.26557 22.26420 19.32375 21.32238 19.32375 19.32375
 [89] 22.41752 20.41889 21.41821 19.47707 23.47707 20.47639 21.47570 19.47707
 [97] 35.90965 28.73648 22.68309 20.83231 18.83368 18.83368 27.68241 32.68172
[105] 27.68241 25.68378 23.68241 26.73785 32.73648 24.73648 25.79603 25.79603
[113] 25.79603 31.79466 19.83299 21.91102 27.90965 24.06297
```

Q9. With the help of a faceted boxplot (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`.
```



Looking at the boxplots, it seems that the two groups are significantly different with aP having less spread around lower ages and wP being more spread out emcompassing higher ages.

#Joining multiple tables

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<-inner_join(specimen,subject)
```

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

```
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
        age
1 13852 days
2 13852 days
3 13852 days
4 13852 days
5 13852 days
6 13852 days
```

Antibody measurments in the blood of patients

```
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
```

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)`
```

```
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 actual_day_relative_to_boost
1 UG/ML                2.096133          1                           -3
2 IU/ML               29.170000          1                           -3
3 IU/ML                0.530000          1                           -3
```

```
4 IU/ML                  6.205949          1                           -3
5 IU/ML                  4.679535          1                           -3
6 IU/ML                  2.816431          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 13852 days
2 13852 days
3 13852 days
4 13852 days
5 13852 days
6 13852 days
```

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 3240 7968 7968 7968 7968
```

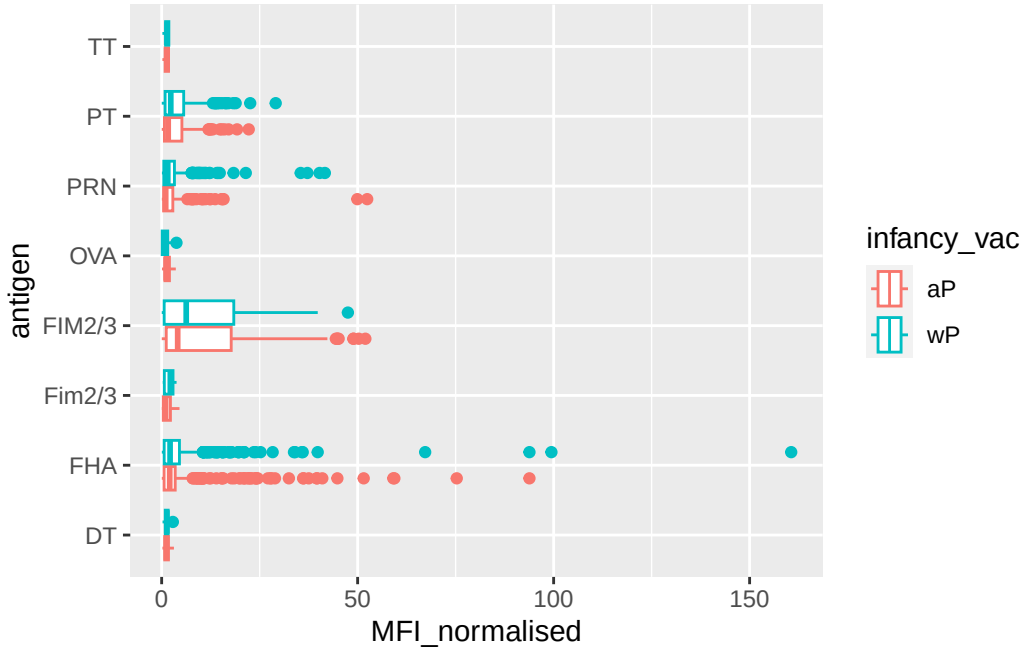
```
igg<-abdata %>% filter (isotype=="IgG")
head(igg)
```

```
  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
         age
1 13852 days
2 13852 days
3 13852 days
4 14948 days
5 14948 days
6 14948 days
```

Boxplot of MFI_normalised vs antigen

```
ggplot(igg)+aes(MFI_normalised, antigen, col=infancy_vac)+geom_boxplot()
```




```
head(igg)
```

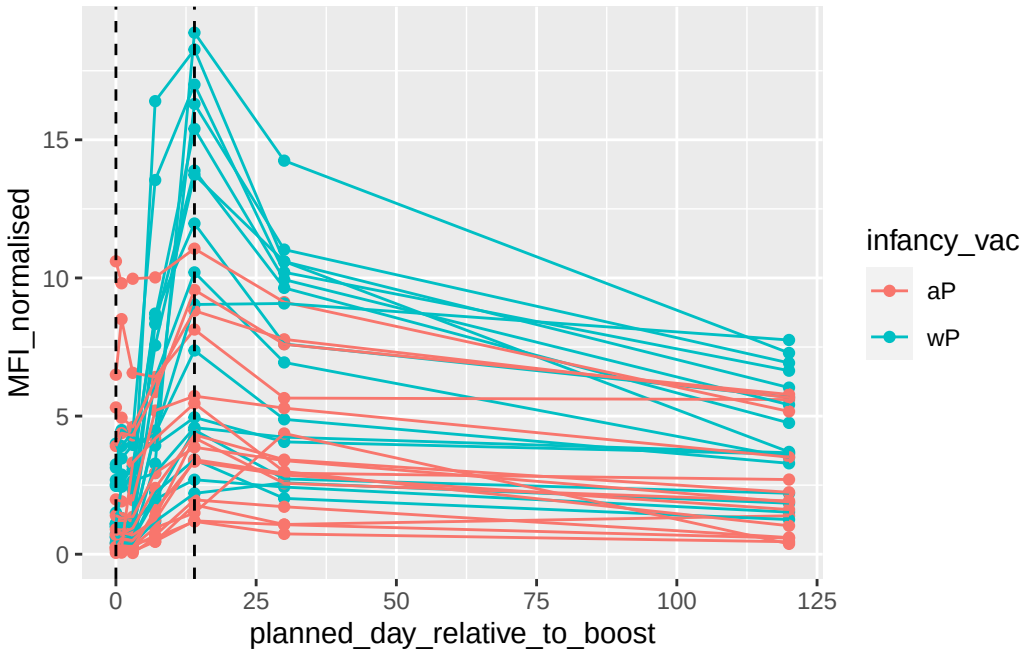
```
  specimen_id isotype is_antigen_specific antigen        MFI MFI_normalised
1           1     IgG                TRUE      PT   68.56614       3.736992
2           1     IgG                TRUE     PRN  332.12718       2.602350
3           1     IgG                TRUE     FHA 1887.12263      34.050956
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
          age
1 13852 days
2 13852 days
3 13852 days
4 14948 days
5 14948 days
6 14948 days
```

Focus on IgG to the Pertussis Toxin (PT) antigen in the 2021 dataset

```
igg.pt<-igg %>% filter(antigen == "PT", dataset=="2021_dataset")
```

```
ggplot(igg.pt)+aes(planned_day_relative_to_boost,MFI_normalised,col=infancy_vac,group=subj
```



Q19.: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

The expression of this gene is significantly higher for the wP infancy vaccine than the aP vaccine at the maximum level suggesting that the evidence that we are getting is starting to show differentiation.