Lab 19

Bomin Xie A16144147

Investigating pertussis cases by year

Q1: (a more complicated way if not using package datapasta, but applicable for any website with xml table)

```
library(xml2)
cdc_table <- read_xml('<table class="table table-bordered table-striped opt-in show-more-d
<caption class="sr-only">table shows reported pertussis cases in the United States since 1

<tbody>
<tr class="expanded">
<th scope="row">1922</th>
<td>107,473</td>
</tr>
<tr class="expanded">
<th scope="row">1923</th>
<td>164,191</td>
</tr>
<tr class="expanded">
<th scope="row">1924</th>
<td>165,418</td>
</tr>
<tr class="expanded">
<th scope="row">1925</th>
<td>152,003</td>
</tr>
<tr class="expanded faded">
<th scope="row">1926</th>
<td>202,210</td>
</tr>
<tr>
```

```
<th scope="row">1927</th>
<td>181,411</td>
</tr>
<tr>
<th scope="row">1928</th>
<td>161,799</td>
</tr>
<tr>
<th scope="row">1929</th>
<td>197,371</td>
</tr>
<tr>
<th scope="row">1930</th>
<td>166,914</td>
</tr>
<tr>
<th scope="row">1931</th>
<td>172,559</td>
</tr>
<tr>
<th scope="row">1932</th>
<td>215,343</td>
</tr>
<tr>
<th scope="row">1933</th>
<td>179,135</td>
</tr>
<tr>
<th scope="row">1934</th>
<td>265,269</td>
</tr>
<tr>
<th scope="row">1935</th>
<td>180,518</td>
</tr>
<tr>
<th scope="row">1936</th>
<td>147,237</td>
</tr>
<tr>
<th scope="row">1937</th>
```

```
<td>214,652</td>
</tr>
<tr>
<th scope="row">1938</th>
<td>227,319</td>
</tr>
<tr>
<th scope="row">1939</th>
<td>103,188</td>
</tr>
<tr>
<th scope="row">1940</th>
<td>183,866</td>
</tr>
<tr>
<th scope="row">1941</th>
<td>222,202</td>
</tr>
<tr>
<th scope="row">1942</th>
<td>191,383</td>
</tr>
<tr>
<th scope="row">1943</th>
<td>191,890</td>
</tr>
<tr>
<th scope="row">1944</th>
<td>109,873</td>
</tr>
<tr>
<th scope="row">1945</th>
<td>133,792</td>
</tr>
<tr>
<th scope="row">1946</th>
<td>109,860</td>
</tr>
<tr>
<th scope="row">1947</th>
<td>156,517</td>
```

```
</tr>
<tr>
<th scope="row">1948</th>
<td>74,715</td>
</tr>
<tr>
<th scope="row">1949</th>
<td>69,479</td>
</tr>
<tr>
<th scope="row">1950</th>
<td>120,718</td>
</tr>
<tr>
<th scope="row">1951</th>
<td>68,687</td>
</tr>
<tr>
<th scope="row">1952</th>
<td>45,030</td>
</tr>
<tr>
<th scope="row">1953</th>
<td>37,129</td>
</tr>
<tr>
<th scope="row">1954</th>
<td>60,886</td>
</tr>
<tr>
<th scope="row">1955</th>
<td>62,786</td>
</tr>
<tr>
<th scope="row">1956</th>
<td>31,732</td>
</tr>
<tr>
<th scope="row">1957</th>
<td>28,295</td>
</tr>
```

```
<tr>
<th scope="row">1958</th>
<td>32,148</td>
</tr>
<tr>
<th scope="row">1959</th>
<td>40,005</td>
</tr>
<tr>
<th scope="row">1960</th>
<td>14,809</td>
</tr>
<tr>
<th scope="row">1961</th>
<td>11,468</td>
</tr>
<tr>
<th scope="row">1962</th>
<td>17,749</td>
</tr>
<tr>
<th scope="row">1963</th>
<td>17,135</td>
</tr>
<tr>
<th scope="row">1964</th>
<td>13,005</td>
</tr>
<tr>
<th scope="row">1965</th>
<td>6,799</td>
</tr>
<tr>
<th scope="row">1966</th>
<td>7,717</td>
</tr>
<tr>
<th scope="row">1967</th>
<td>9,718</td>
</tr>
<tr>
```

```
<th scope="row">1968</th>
<td>4,810</td>
</tr>
<tr>
<th scope="row">1969</th>
<td>3,285</td>
</tr>
<tr>
<th scope="row">1970</th>
<td>4,249</td>
</tr>
<tr>
<th scope="row">1971</th>
<td>3,036</td>
</tr>
<tr>
<th scope="row">1972</th>
<td>3,287</td>
</tr>
<tr>
<th scope="row">1973</th>
<td>1,759</td>
</tr>
<tr>
<th scope="row">1974</th>
<td>2,402</td>
</tr>
<tr>
<th scope="row">1975</th>
<td>1,738</td>
</tr>
<tr>
<th scope="row">1976</th>
<td>1,010</td>
</tr>
<tr>
<th scope="row">1977</th>
<td>2,177</td>
</tr>
<tr>
<th scope="row">1978</th>
```

```
<td>2,063</td>
</tr>
<tr>
<th scope="row">1979</th>
<td>1,623</td>
</tr>
<tr>
<th scope="row">1980</th>
<td>1,730</td>
</tr>
<tr>
<th scope="row">1981</th>
<td>1,248</td>
</tr>
<tr>
<th scope="row">1982</th>
<td>1,895</td>
</tr>
<tr>
<th scope="row">1983</th>
<td>2,463</td>
</tr>
<tr>
<th scope="row">1984</th>
<td>2,276</td>
</tr>
<tr>
<th scope="row">1985</th>
<td>3,589</td>
</tr>
<tr>
<th scope="row">1986</th>
<td>4,195</td>
</tr>
<tr>
<th scope="row">1987</th>
<td>2,823</td>
</tr>
<tr>
<th scope="row">1988</th>
<td>3,450</td>
```

```
</tr>
<tr>
<th scope="row">1989</th>
<td>4,157</td>
</tr>
<tr>
<th scope="row">1990</th>
<td>4,570</td>
</tr>
<tr>
<th scope="row">1991</th>
<td>2,719</td>
</tr>
<tr>
<th scope="row">1992</th>
<td>4,083</td>
</tr>
<tr>
<th scope="row">1993</th>
<td>6,586</td>
</tr>
<tr>
<th scope="row">1994</th>
<td>4,617</td>
</tr>
<tr>
<th scope="row">1995</th>
<td>5,137</td>
</tr>
<tr>
<th scope="row">1996</th>
<td>7,796</td>
</tr>
<tr>
<th scope="row">1997</th>
<td>6,564</td>
</tr>
<tr>
<th scope="row">1998</th>
<td>7,405</td>
</tr>
```

```
<tr>
<th scope="row">1999</th>
<td>7,298</td>
</tr>
<tr>
<th scope="row">2000</th>
<td>7,867</td>
</tr>
<tr>
<th scope="row">2001</th>
<td>7,580</td>
</tr>
<tr>
<th scope="row">2002</th>
<td>9,771</td>
</tr>
<tr>
<th scope="row">2003</th>
<td>11,647</td>
</tr>
<tr>
<th scope="row">2004</th>
<td>25,827</td>
</tr>
<tr>
<th scope="row">2005</th>
<td>25,616</td>
</tr>
<tr>
<th scope="row">2006</th>
<td>15,632</td>
</tr>
<tr>
<th scope="row">2007</th>
<td>10,454</td>
</tr>
<tr>
<th scope="row">2008</th>
<td>13,278</td>
</tr>
<tr>
```

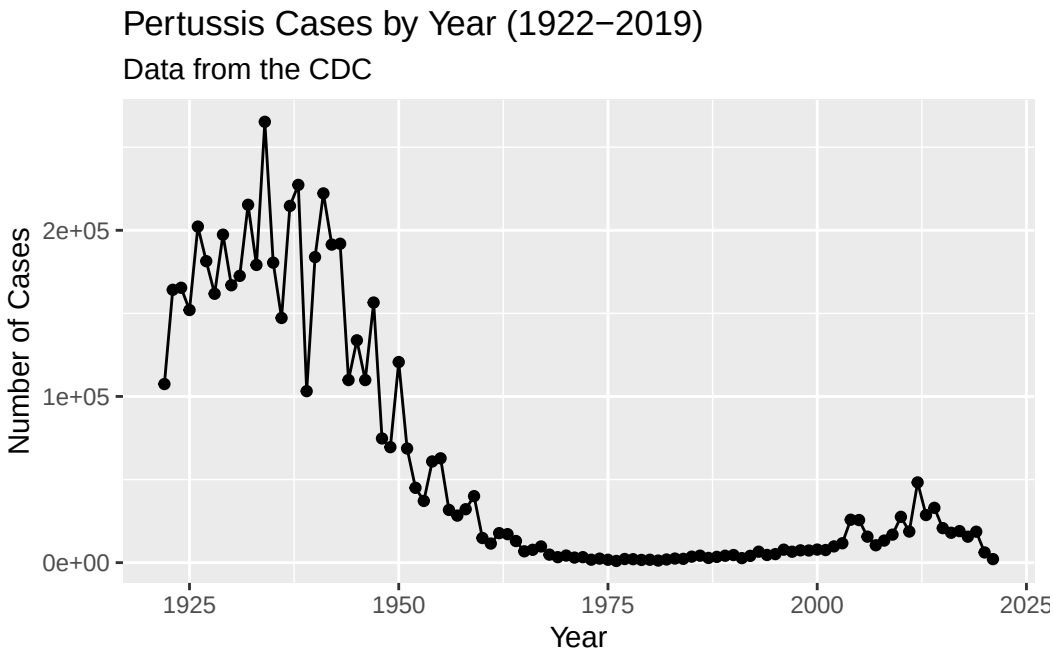
```
<th scope="row">2009</th>
<td>16,858</td>
</tr>
<tr>
<th scope="row">2010</th>
<td>27,550</td>
</tr>
<tr>
<th scope="row">2011</th>
<td>18,719</td>
</tr>
<tr>
<th scope="row">2012</th>
<td>48,277</td>
</tr>
<tr>
<th scope="row">2013</th>
<td>28,639</td>
</tr>
<tr>
<th scope="row">2014</th>
<td>32,971</td>
</tr>
<tr>
<th scope="row">2015</th>
<td>20,762</td>
</tr>
<tr>
<th scope="row">2016</th>
<td>17,972</td>
</tr>
<tr>
<th scope="row">2017</th>
<td>18,975</td>
</tr>
<tr>
<th scope="row">2018</th>
<td>15,609</td>
</tr>
<tr>
<th scope="row">2019</th>
```

```
<td>18,617</td>
</tr>
<tr>
<th scope="row">2020</th>
<td>6,124</td>
</tr>
<tr>
<th scope="row">2021</th>
<td>2,116</td>
</tr>
</tbody>
</table>') # import the data from CDC website
```

```
pertussis_year <- xml_text(xml_find_all(cdc_table, xpath = "//th"))
pertussis_data <- xml_text(xml_find_all(cdc_table, xpath = "//td"))
cdc <- data.frame(Year = as.numeric(pertussis_year), Cases = as.numeric(gsub(",","",pertus
head(cdc)
```

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

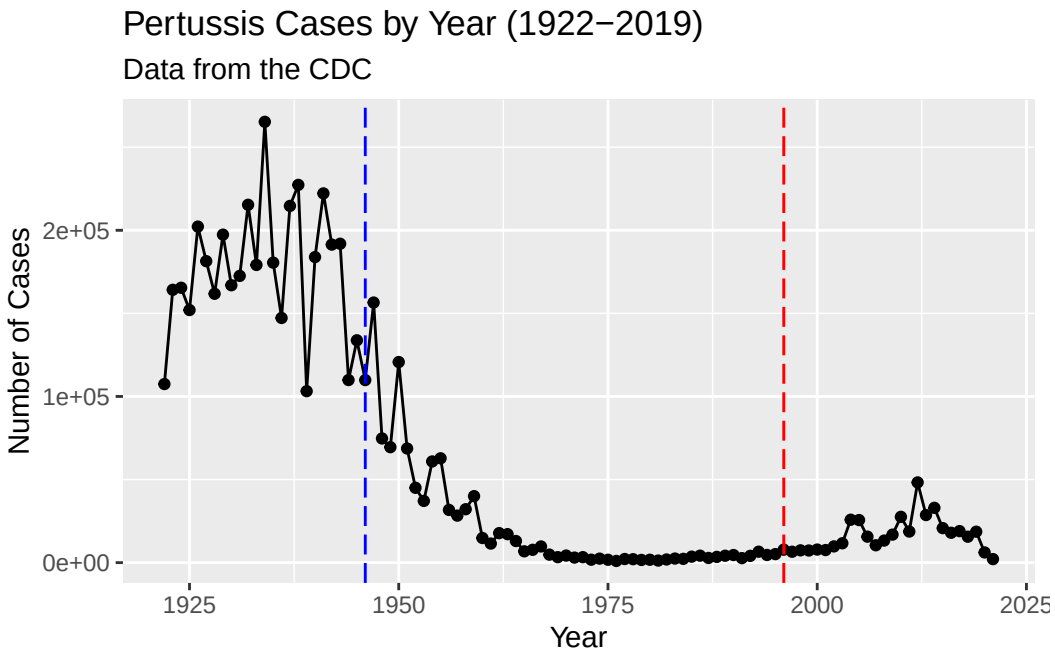
```
library(ggplot2)
base <- ggplot(cdc) +
  aes(Year, Cases) +
  geom_point() +
  geom_line() +
  labs(x= "Year", y= "Number of Cases", title = "Pertussis Cases by Year (1922-2019)", sub
base
```



A tale of two vaccines

Q2: Based on the figure below, the introduction of aP vaccine has a lower number of cases.

```
base + geom_vline(xintercept= 1946, col = "blue", linetype=5) +
  geom_vline(xintercept = 1996, col= "red", linetype= 5)
```



Q3: after the introduction of aP vaccine, the number of Pertussis remain low which is different from the wP vaccine. The uprising of cases might be contributed to the anti-vaccine activities across the world.

Exploring CMI-PB data

```
library(jsonlite)

subject <- read_json("http://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:

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

```
aP wP
47 49
```

Based on the above result, 47 aP infants and 49 wP infants are in the dataset.

Q5:

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

```
Female   Male
    66     30
```

Based on the result above, 66 femal and 30 male are in the dataset.

Q6:

```
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
```

The breakdown of race and biological sex are listed above.

Working with dates

Q7:

```
library(lubridate)
```

```
Attaching package: 'lubridate'
```

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

    date, intersect, setdiff, union
```

```
# Use todays date to calculate age in days
subject$age <- today() - ymd(subject$year_of_birth)
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        age
1    1986-01-01    2016-09-12 2020_dataset 13677 days
2    1968-01-01    2019-01-28 2020_dataset 20252 days
3    1983-01-01    2016-10-10 2020_dataset 14773 days
4    1988-01-01    2016-08-29 2020_dataset 12947 days
5    1991-01-01    2016-08-29 2020_dataset 11851 days
6    1988-01-01    2016-10-10 2020_dataset 12947 days
```

```
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      26      26      27 
```

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

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
     28      32      35      37      40      55 
```

```r
t.test(ap$age, wp$age)
```

```
    Welch Two Sample t-test

data:  ap$age and wp$age
t = -12.092 days, df = 51.082, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4618.534 days -3303.337 days
sample estimates:
Time differences in days
mean of x mean of y 
 9410.574 13371.510 
```

Therefore, the average age of wP individuals are 37, and aP individuals are 26. T-test shows the two groups are significantly different.

Q8:

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

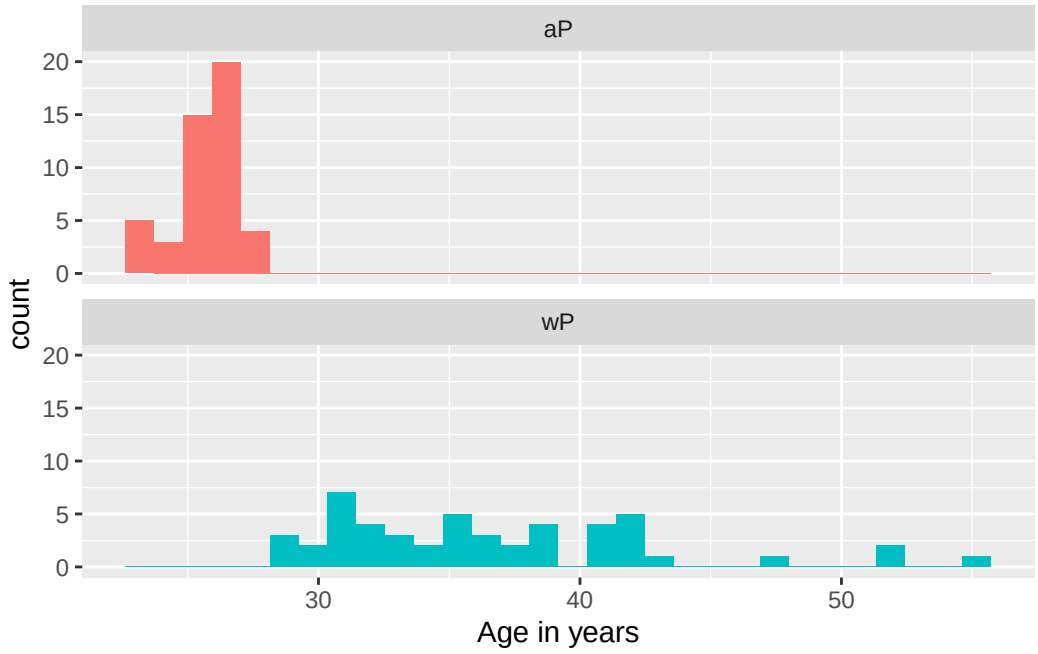
```
head(age_at_boost)
```

```
[1] 30.69678 51.07461 33.77413 28.65982 25.65914 28.77481
```

Q9:

```
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`.
```



Joining multiple tables

```
# Complete the API URLs...
specimen <- read_json("http://cmi-pb.org/api/specimen", simplifyVector = TRUE)
titer <- read_json("http://cmi-pb.org/api/ab_titer", simplifyVector = TRUE)
```

Q10:

```
meta <- inner_join(specimen, subject)
```

```
Joining with `by = join_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 13677 days
2 13677 days
3 13677 days
4 13677 days
5 13677 days
6 13677 days
```

Q11:

```
abdata <- inner_join(titer, meta)
```

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

```
dim(abdata)
```

```
[1] 32675    21
```

Q12:

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

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

Based on the above response, there are more than 1000 specimens for each isotype.

Q13:

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

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

The vist 8 specimens are much less than the others.

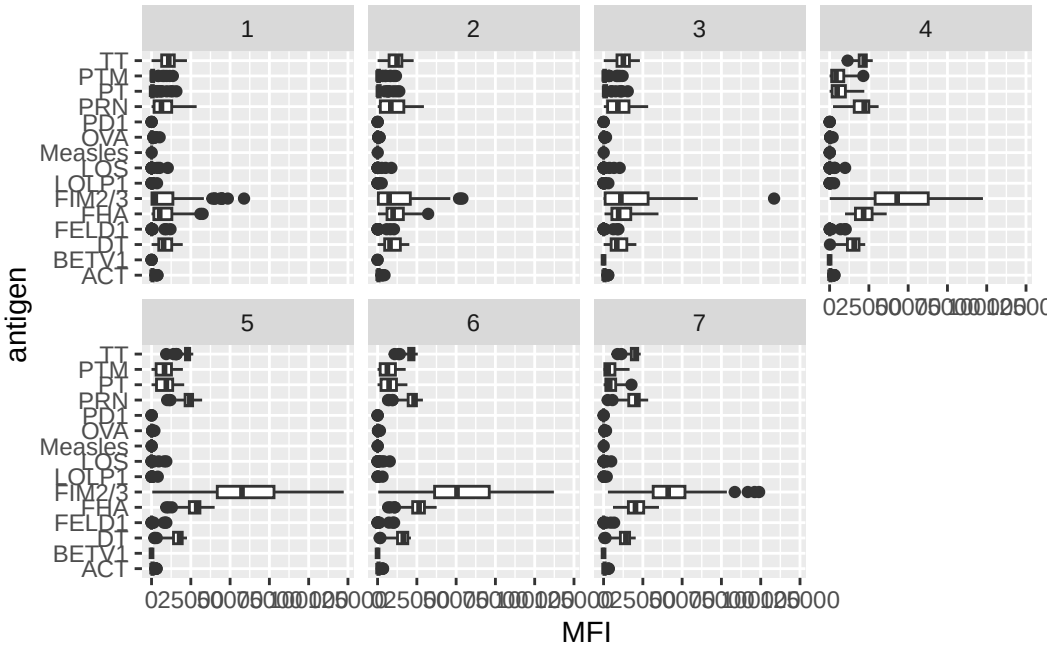
Exmaine IgG1 Ab titer levels

```
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 13677 days
2 13677 days
3 13677 days
4 13677 days
5 13677 days
6 13677 days
```

Q14:

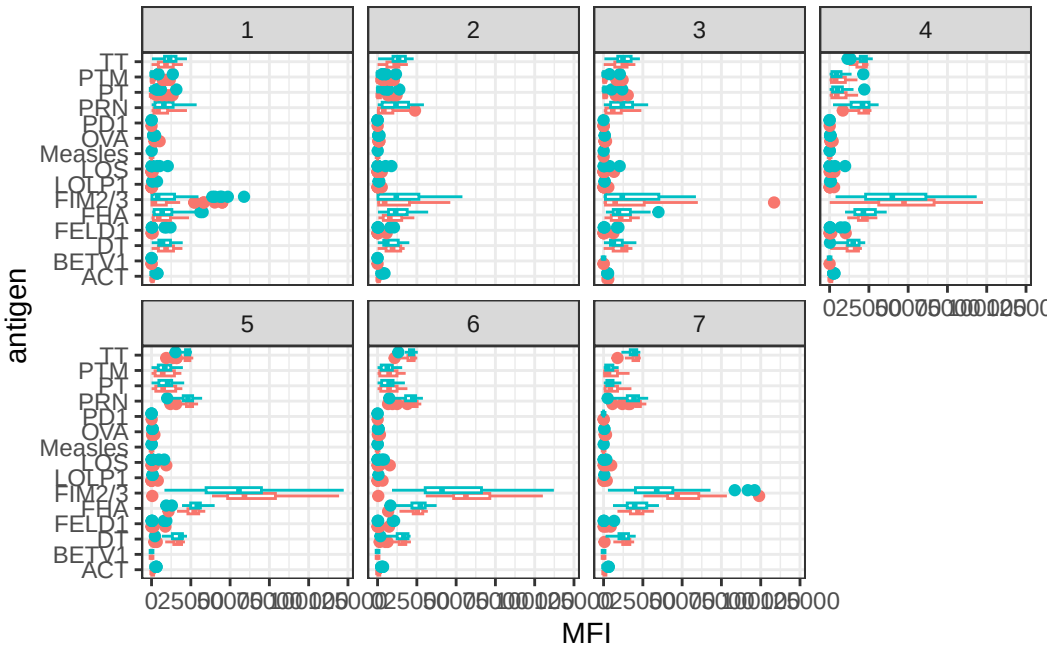
```
ggplot(ig1) +
  aes(MFI, antigen) +
  geom_boxplot() +
  facet_wrap(vars(visit), nrow=2)
```



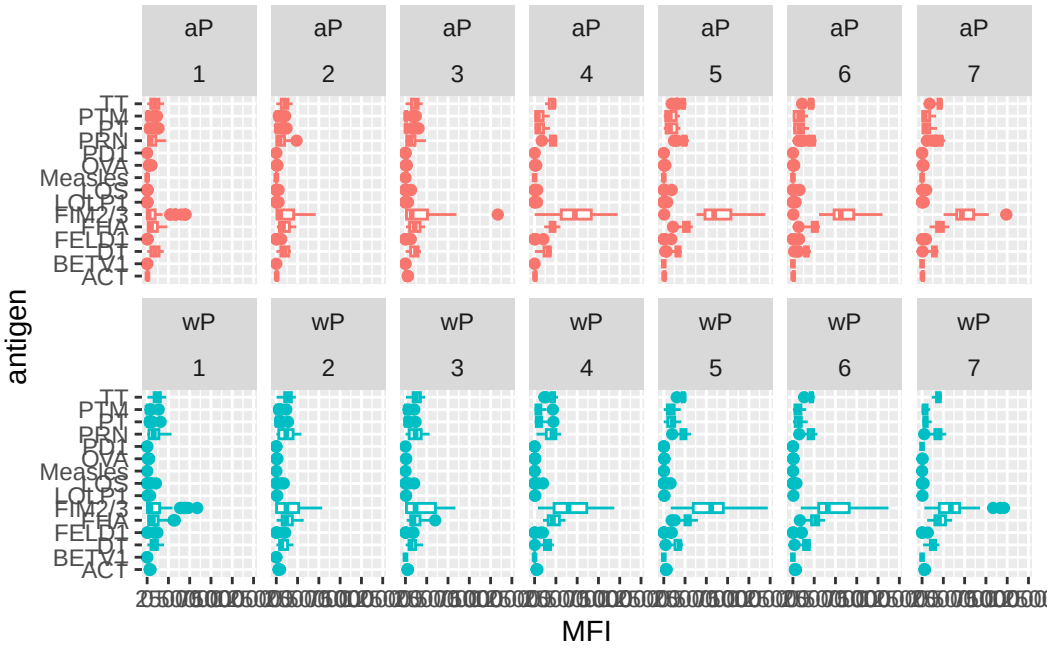
Q15:

Based on the above graph, FIM2/3 shows the most difference in levels of IgG1 antibodies. FIM2/3 protein is presented in the aP and wP vaccine, which has the function in cell adhesion and infection locating on the outside of pathogen. Therefore, other antigens are not much involved in the pertussis infection causing the less antibodies against them.

```
ggplot(ig1) +
  aes(MFI, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  theme_bw()
```

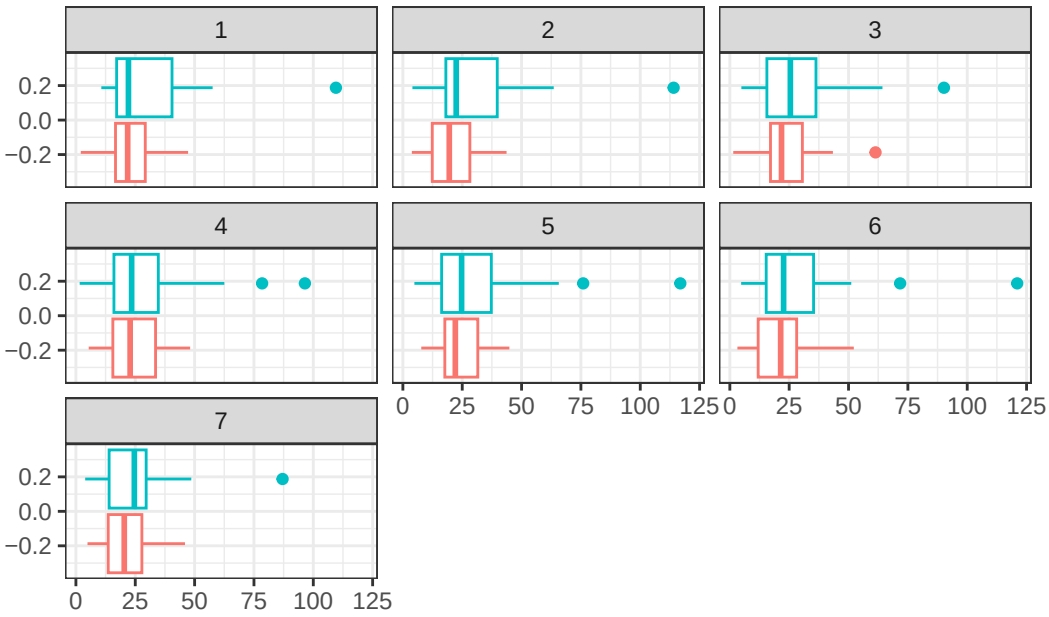



```
ggplot(ig1) +
  aes(MFI, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)
```

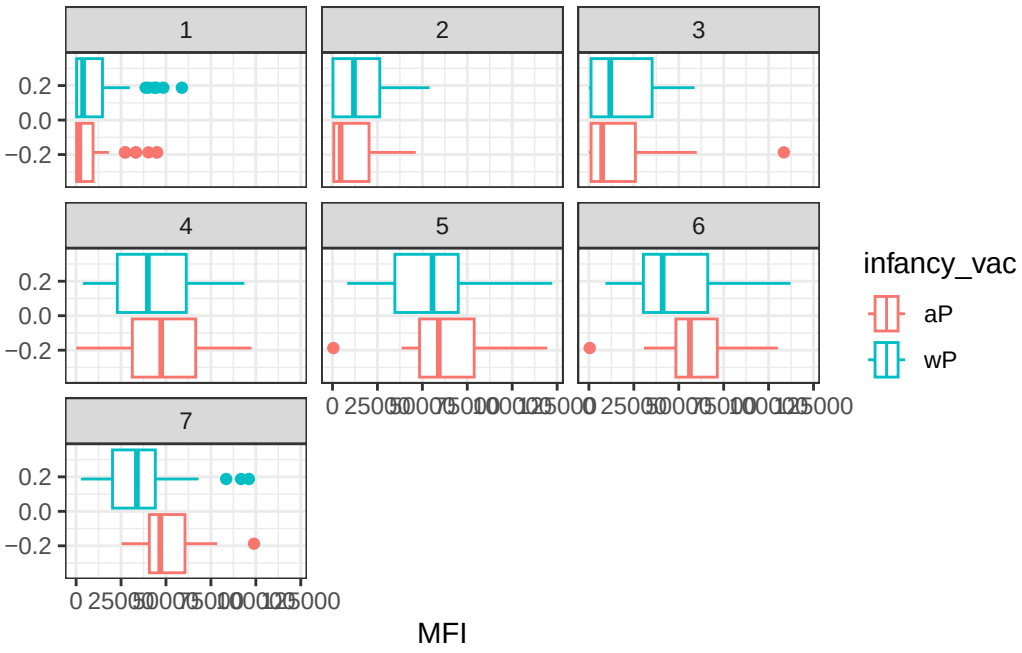


Q16:

```
filter(ig1, antigen=="Measles") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```




```
filter(ig1, antigen== "FIM2/3") %>%
  ggplot() +
  aes(MFI, col=infancy_vac) +
  geom_boxplot(show.legend = TRUE) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



Q17: When examining the FIM2/3 antigen, one significant finding is that the median MFI level during the initial visit is extremely low, approximately 0. There is a progressive rise in antibody levels to approximately MFI: 50000.

Q18: In terms of the FIM2/3 antigen, it appears that the aP vaccine demonstrates a more gradual increase in antibody levels. The wP vaccine exhibits a rapid increase in antibody levels, particularly after the third visit. The MFI levels of the aP vaccine reach a higher point compared to the MFI levels of the wP vaccine.

Obtaining CMI-PB RNASeq data

```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.
rna <- read_json(url, simplifyVector = TRUE)
meta <- inner_join(specimen, subject)
```

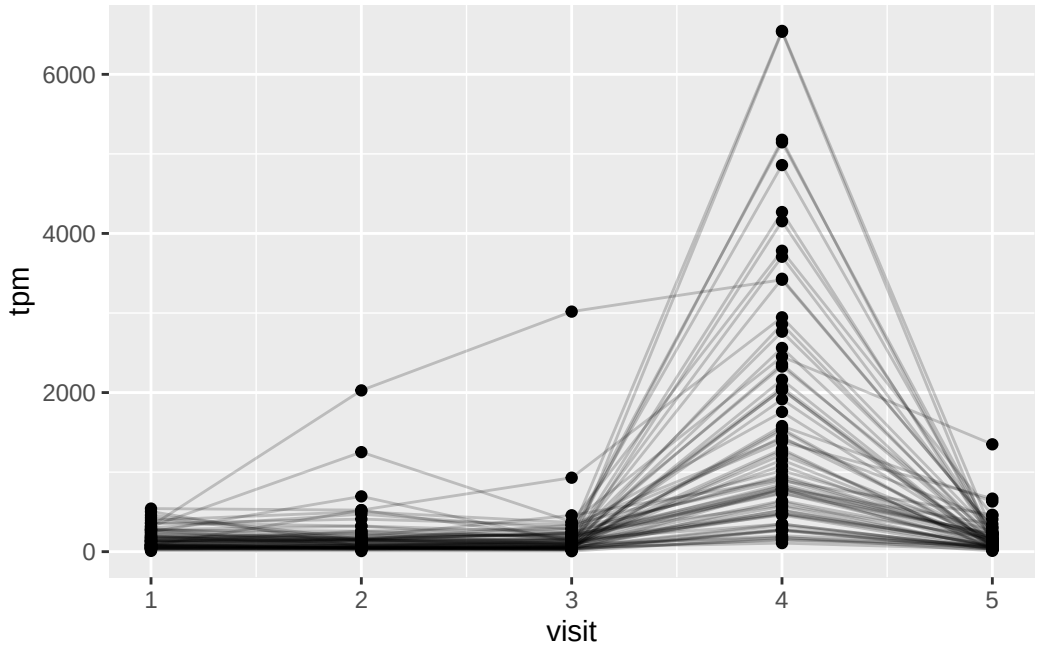
```
Joining with `by = join_by(subject_id)`
```

```
ssrna <- inner_join(rna, meta)
```

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

Q19:

```
ggplot(ssrna) +
  aes(visit, tpm, group=subject_id) +
  geom_point() +
  geom_line(alpha=0.2)
```

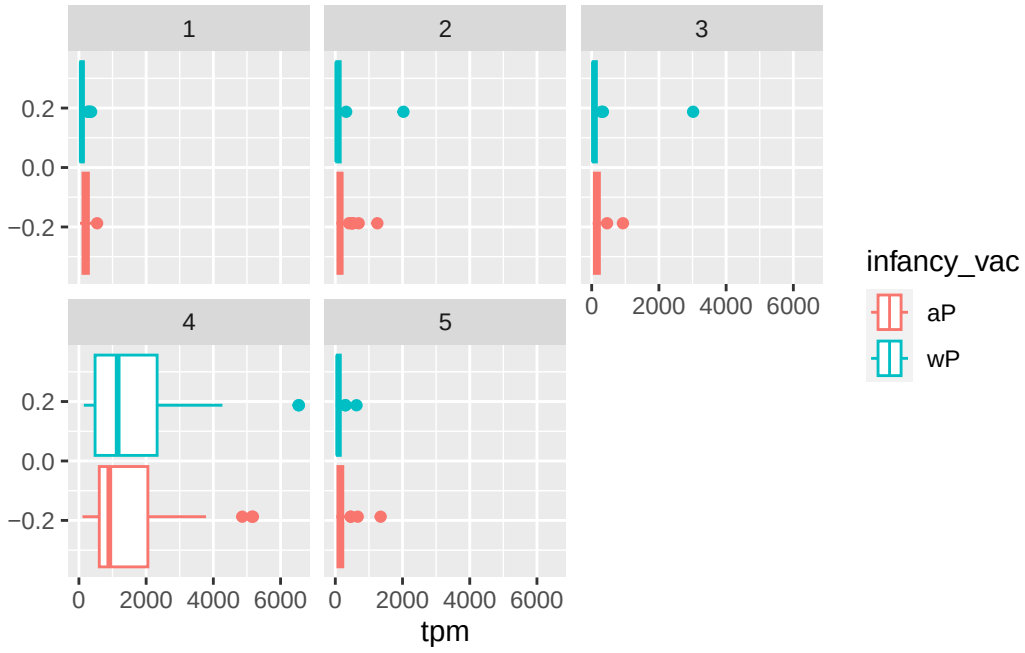


Q20: This gene reaches its highest expression level during visit 4. The tpm metric measures the abundance of transcripts per million, and at visit 4, the gene demonstrates its maximum expression level.

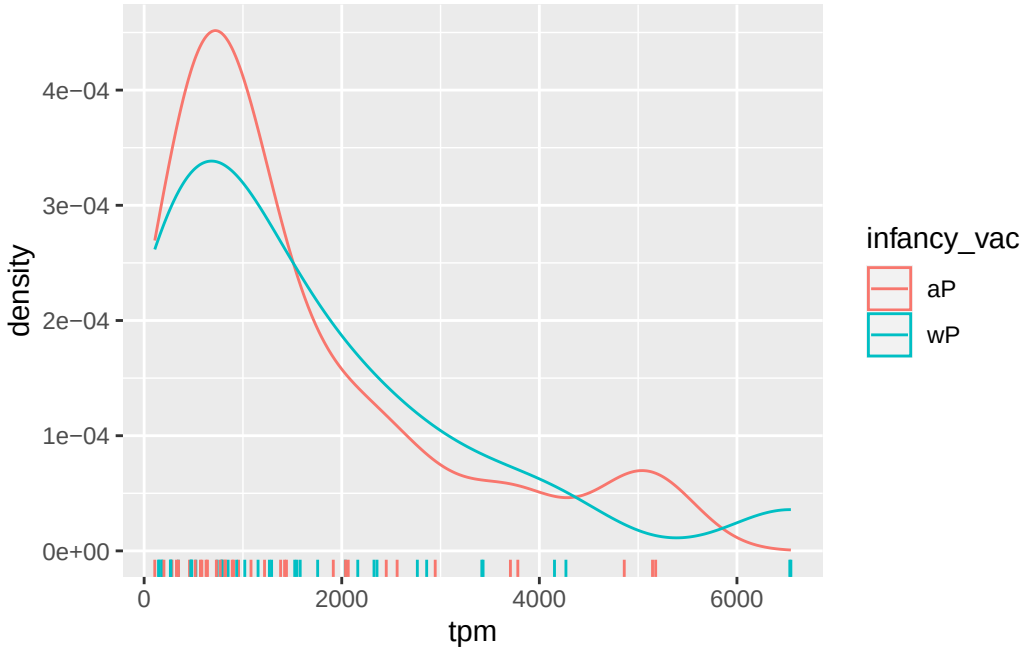
Q21: The gene involved in immunoglobulin construction would exhibit its maximum expression level around visit 4. This aligns with the observation that the antibody titer, which is an indicator of the immune response, reaches its peak level around the same visit. The gene's heightened expression suggests its crucial role in the production of immunoglobulins, contributing to the increased antibody levels observed during that period.

```
ggplot(ssrna) +
  aes(tpm, col=infancy_vac) +
  geom_boxplot() +
```

```
  facet_wrap(vars(visit))
```




```
ssrna %>%
  filter(visit==4) %>%
  ggplot() +
    aes(tpm, col=infancy_vac) + geom_density() +
    geom_rug()
```


density
4e−04
3e−04
2e−04
1e−04
0e+00
0
2000
4000
6000
tpm
infancy_vac
aP
wP