Class19

Brian Wells (PID: A69026838)

#investigate pertussis cases by year

Here is where the numbers are from: https://www.cdc.gov/pertussis/surv-reporting/cases-by-year.html

> 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.

```
#install.packages("datapasta")
library(datapasta)
library(tidyverse)
```

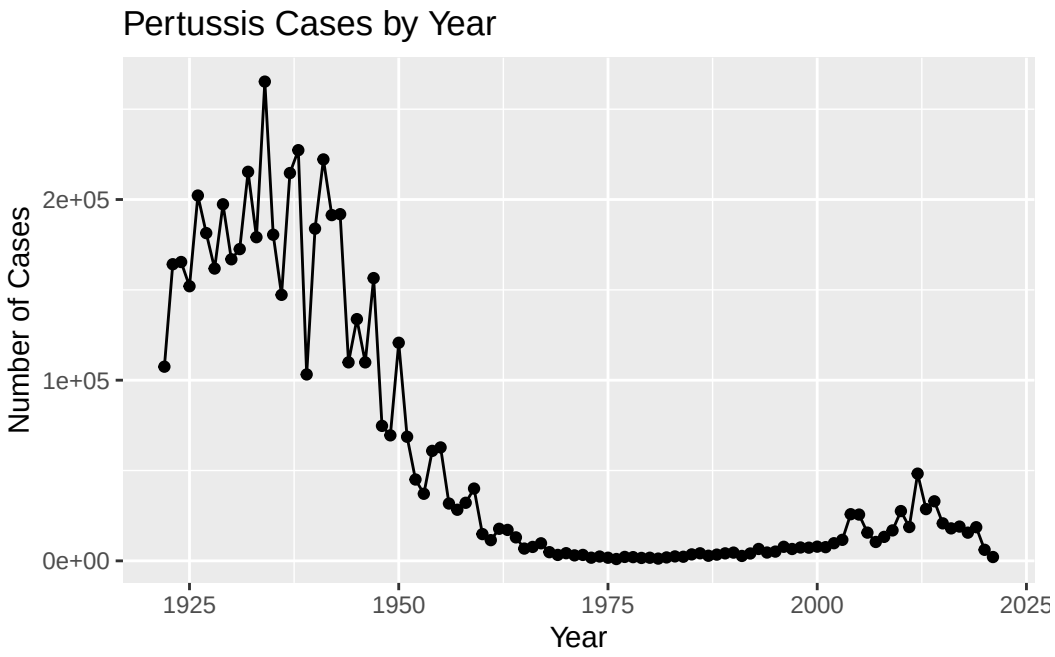
```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.4.4     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
#| 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),
  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)
)
```

Here's the plot

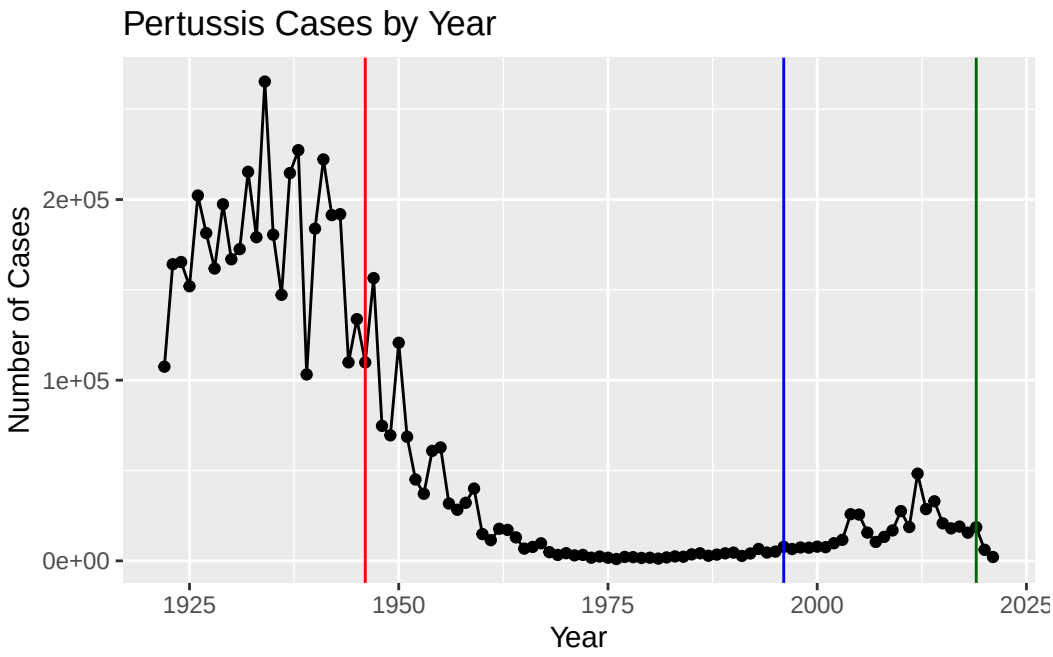
```
ggplot(cdc) +
  aes(Year, No.Reported.Pertussis.Cases) +
  geom_point()+
  geom_line() +
  labs(title="Pertussis Cases by Year", x="Year", y="Number of Cases")
```



#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(Year, No.Reported.Pertussis.Cases) +
  geom_point()+
  geom_line() +
  geom_vline(xintercept=1946, color="red") +
  geom_vline(xintercept=1996, color="blue")+
  geom_vline(xintercept=2019, color="darkgreen")+
  labs(title="Pertussis Cases by Year", x="Year", y="Number of Cases")
```



at 1946, vaccination use brought cases down to nearly zero. at year 1996, the numbers start to rise.

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

after aP vax introduction, numbers begin to rise. This may be due to waning immunity over time. It could also be that the aP vax. It could also be that the tests got more sensitive. Additionally, there was an antivax movement at the time.

#The CMI-PB API returns JSON data

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

```
Attaching package: 'jsonlite'
```

```
The following object is masked from 'package:purrr':

    flatten
```

```
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
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
```

#Side-Note: Working with dates

```
library(lubridate)
#What is today's date
today()
```

```
[1] "2023-12-06"
```

```
#How many days have passed since new year 2000
today() - ymd("2000-01-01")
```

```
Time difference of 8740 days
```

```
#What is this in years?
time_length( today() - ymd("2000-01-01"),  "years")
```

```
[1] 23.92882
```

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)

ap <- subject %>% filter(infancy_vac == "aP")

summary( time_length( ap$age, "years" ) )
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.93   25.93   25.93   26.03   26.93   29.93
```

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

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  27.93   31.18   35.43   36.32   38.93   55.93
```

```
tt <- t.test(time_length(wp$age, "years"),
             time_length(ap$age, "years"))
tt$p.value
```

```
[1] 6.813505e-19
```

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

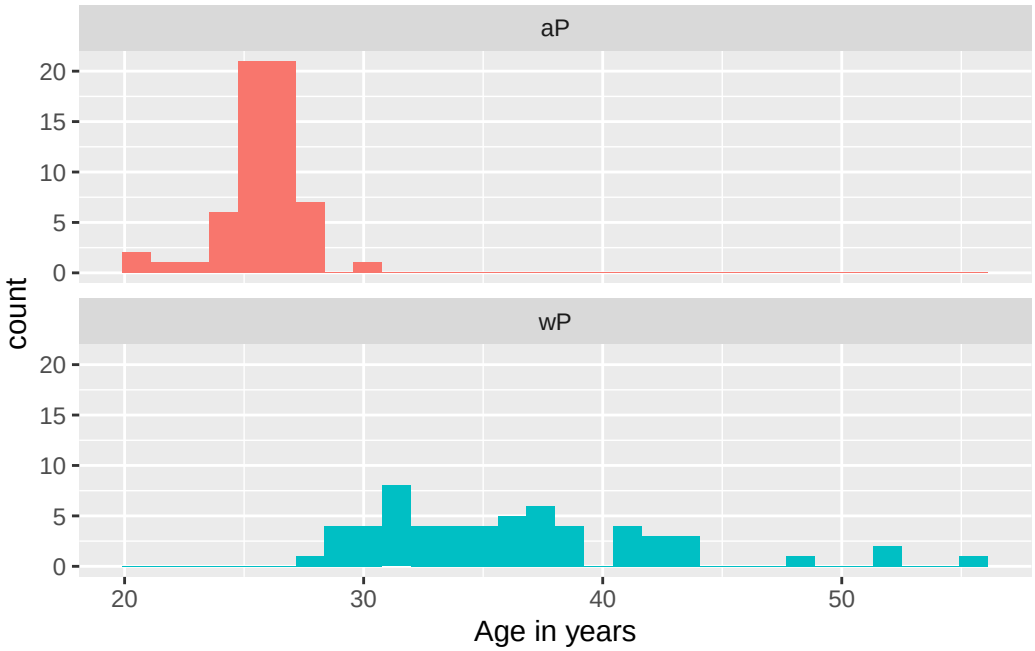
```
int <- ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
age_at_boost <- time_length(int, "year")
head(age_at_boost)
```

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

Q9. 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) +
  xlab("Age in years")
```

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



#Joining multiple tables

```
specimen <- read_json("https://www.cmi-pb.org/api/specimen", simplifyVector = T)
titer <- read_json("https://www.cmi-pb.org/api/v4/plasma_ab_titer", simplifyVector = T)
```

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)`
```

```
dim(meta)
```

```
[1] 939  14
```

```
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 13853 days
2 13853 days
3 13853 days
4 13853 days
5 13853 days
6 13853 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 with `by = join_by(specimen_id)`
```

```
dim(abdata)
```

```
[1] 41810    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 3240 7968 7968 7968 7968
```

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

```
table(abdata$dataset)
```

```
2020_dataset 2021_dataset 2022_dataset
       31520         8085         2205
```

#4. Examine IgG Ab titer levels

```
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 13853 days
2 13853 days
3 13853 days
4 14949 days
5 14949 days
6 14949 days
```

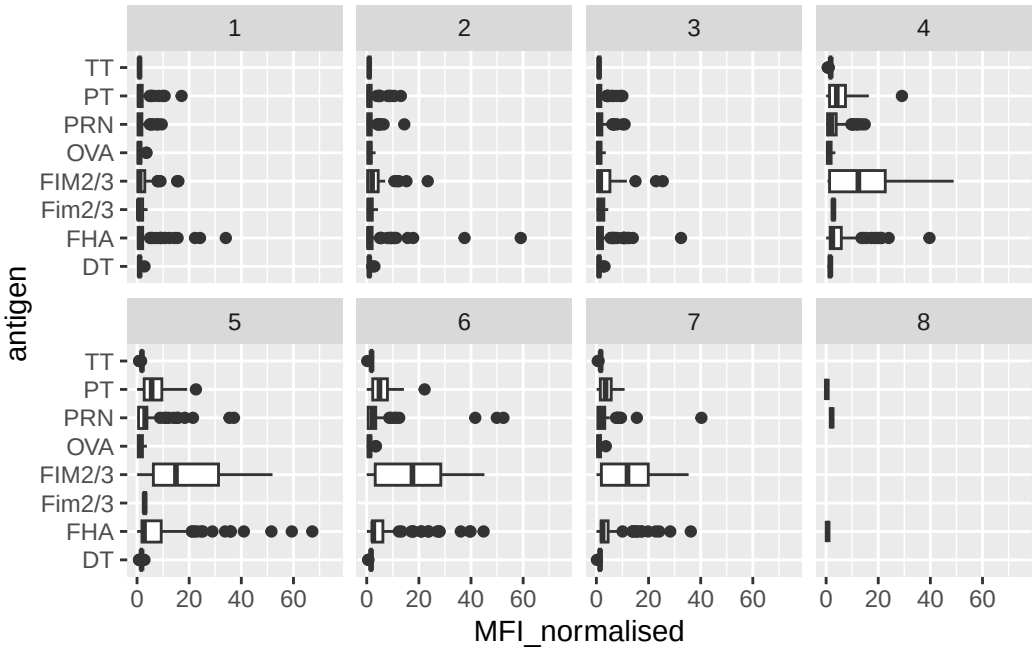
```
colnames(igg)
```

```
 [1] "specimen_id"                    "isotype"
 [3] "is_antigen_specific"            "antigen"
 [5] "MFI"                            "MFI_normalised"
 [7] "unit"                           "lower_limit_of_detection"
 [9] "subject_id"                     "actual_day_relative_to_boost"
[11] "planned_day_relative_to_boost" "specimen_type"
[13] "visit"                          "infancy_vac"
[15] "biological_sex"                 "ethnicity"
[17] "race"                           "year_of_birth"
[19] "date_of_boost"                  "dataset"
[21] "age"
```

Q13. Complete the following code to make a summary boxplot of Ab titer levels (MFI) for all antigens:

```
ggplot(igg) +
  aes(MFI_normalised, antigen) +
  geom_boxplot() +
    xlim(0,75) +
  facet_wrap(vars(visit), nrow=2)
```

```
Warning: Removed 5 rows containing non-finite values (`stat_boxplot()`).
```

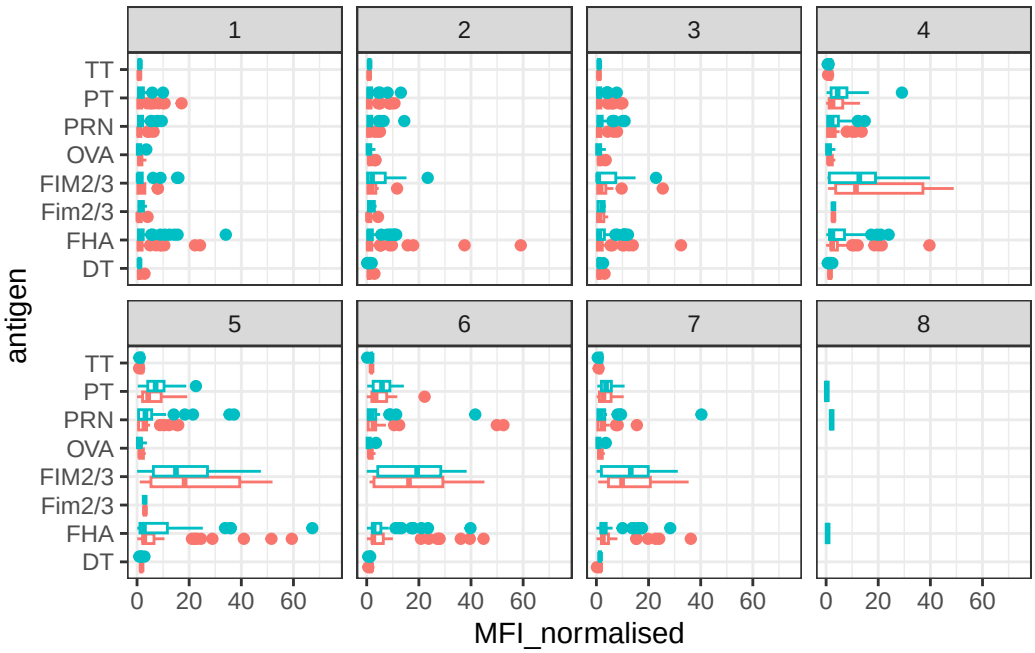


Q14. What antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others?

FIM2/3, PT, FHA, PRN. They're associated with pertussis and are present in the aP vaccine.

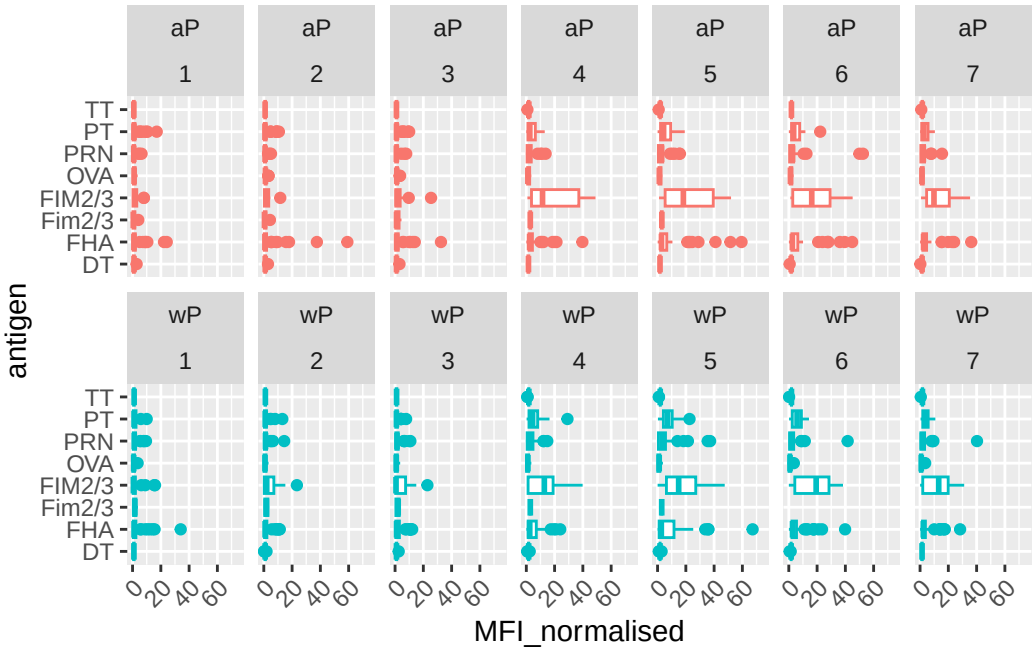
```
ggplot(igg) +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  facet_wrap(vars(visit), nrow=2) +
  xlim(0,75) +
  theme_bw()
```

```
Warning: Removed 5 rows containing non-finite values (`stat_boxplot()`).
```



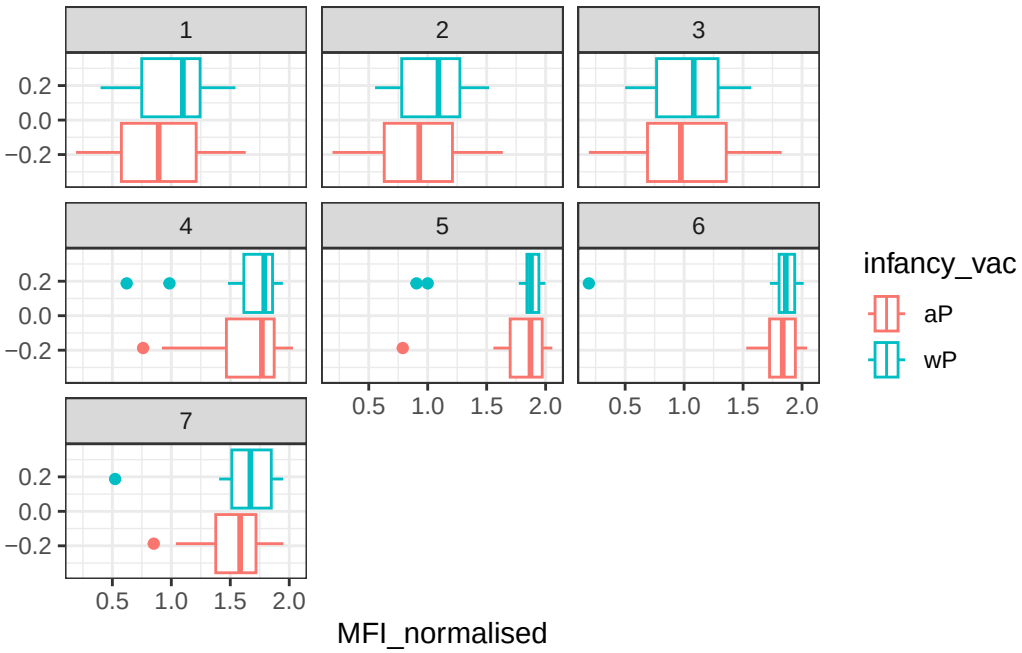

```
igg %>% filter(visit != 8) %>%
ggplot() +
  aes(MFI_normalised, antigen, col=infancy_vac ) +
  geom_boxplot(show.legend = FALSE) +
  xlim(0,75) +
  facet_wrap(vars(infancy_vac, visit), nrow=2)+
  theme(axis.text.x = element_text(angle = 45, hjust=1))
```

```
Warning: Removed 5 rows containing non-finite values (`stat_boxplot()`).
```

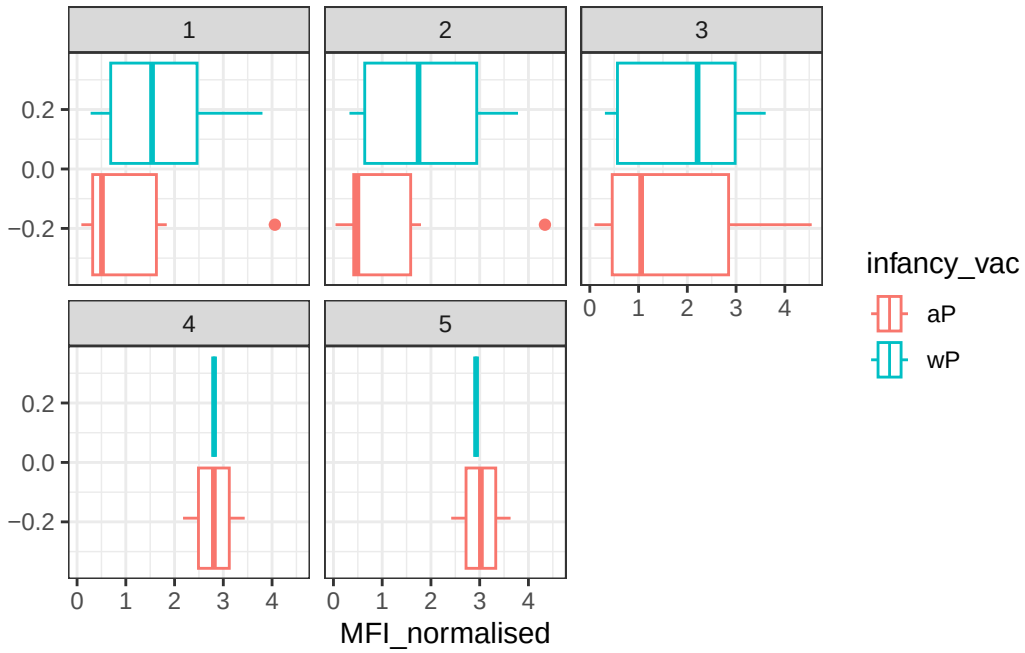


Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("OVA", that is not in our vaccines) and a clear antigen of interest ("PT", Pertussis Toxin, one of the key virulence factors produced by the bacterium B. pertussis).

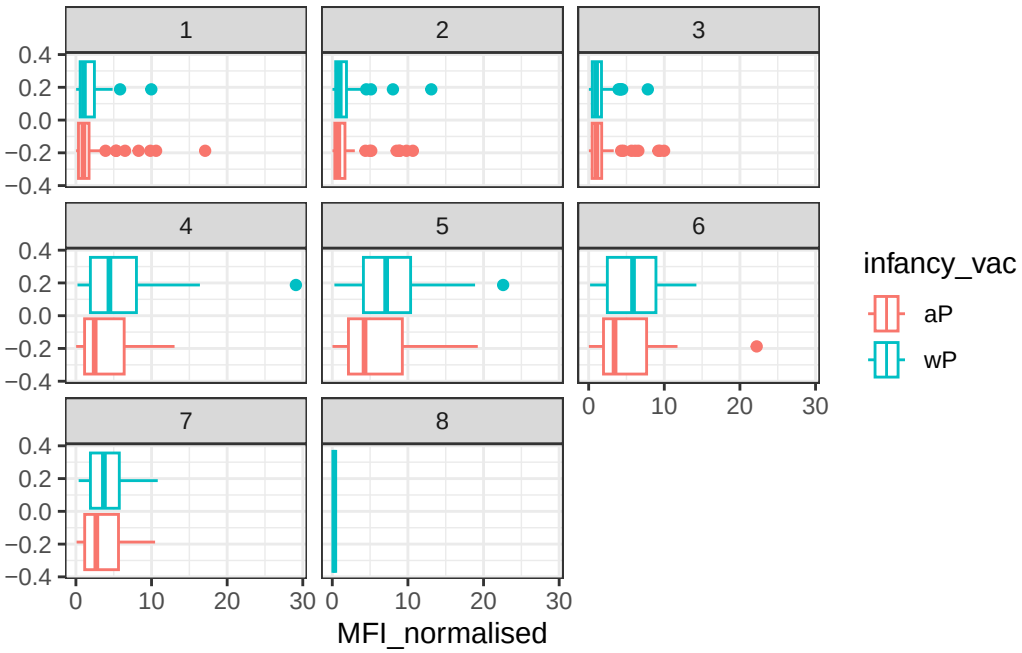
```
filter(igg, antigen=="TT") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = T) +
  facet_wrap(vars(visit)) +
  theme_bw()
```




```
filter(igg, antigen=="Fim2/3") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = T) +
  facet_wrap(vars(visit)) +
  theme_bw()
```




```
filter(igg, antigen=="PT") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = T) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



Q16. What do you notice about these two antigens time courses and the PT data in particular?

The PT levels seem to rise over time and are WAY more than TT.

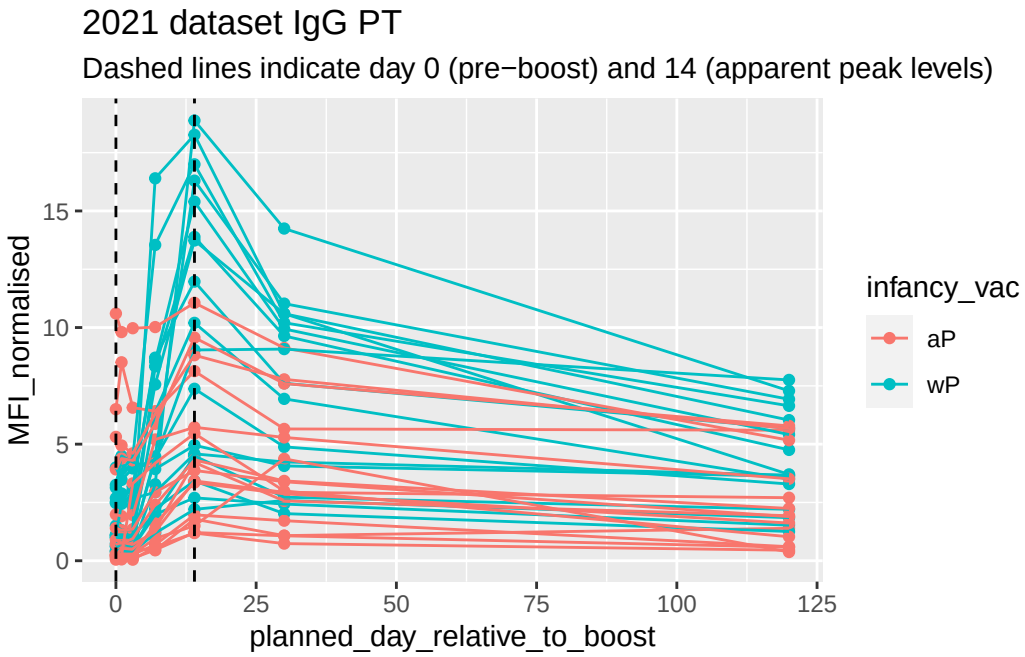
Q17. Do you see any clear difference in aP vs. wP responses?

There doesn't appear to be mnuch of a difference between aP and wP responses.

```
abdata.21 <- abdata %>% filter(dataset == "2021_dataset")

abdata.21 %>%
  filter(isotype == "IgG",  antigen == "PT") %>%
  ggplot() +
    aes(x=planned_day_relative_to_boost,
        y=MFI_normalised,
        col=infancy_vac,
        group=subject_id) +
    geom_point() +
    geom_line() +
    geom_vline(xintercept=0, linetype="dashed") +
    geom_vline(xintercept=14, linetype="dashed") +
  labs(title="2021 dataset IgG PT",
```

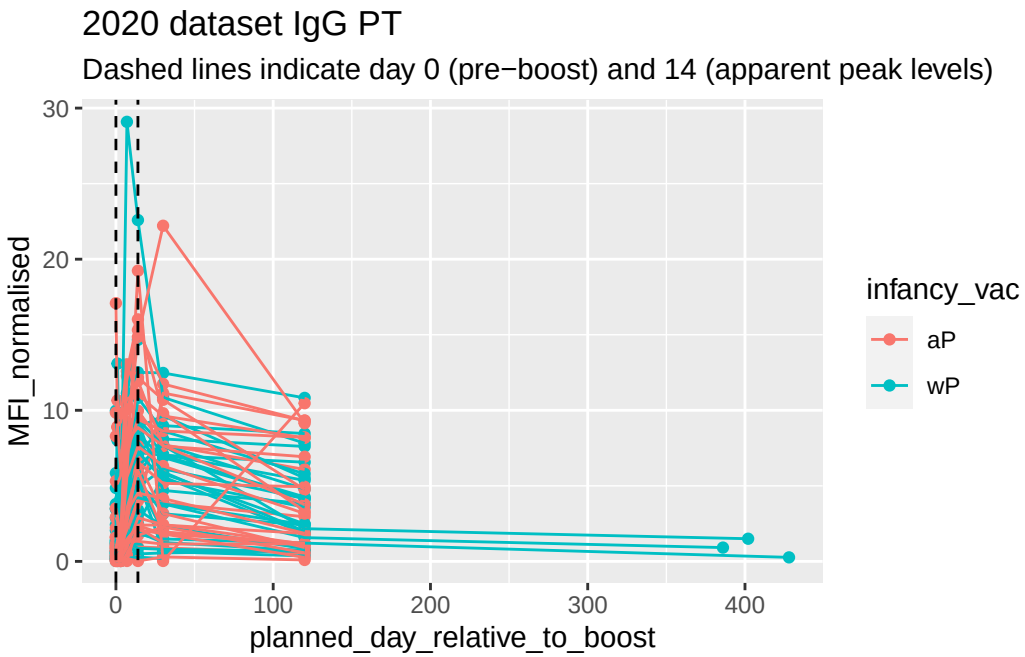
```
      subtitle = "Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)")
```



Q18. Does this trend look similar for the 2020 dataset?

```
abdata.20 <- abdata %>% filter(dataset == "2020_dataset")

abdata.20 %>%
  filter(isotype == "IgG",  antigen == "PT") %>%
  ggplot() +
    aes(x=planned_day_relative_to_boost,
        y=MFI_normalised,
        col=infancy_vac,
        group=subject_id) +
    geom_point() +
    geom_line() +
    geom_vline(xintercept=0, linetype="dashed") +
    geom_vline(xintercept=14, linetype="dashed") +
  labs(title="2020 dataset IgG PT",
       subtitle = "Dashed lines indicate day 0 (pre-boost) and 14 (apparent peak levels)")
```



Woah. COVID clearly messed up this data set. generally the trend is similar, considering that the peak coincides with the d14 vertical line. There

#5. Obtaining CMI-PB RNASeq data

```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.
```

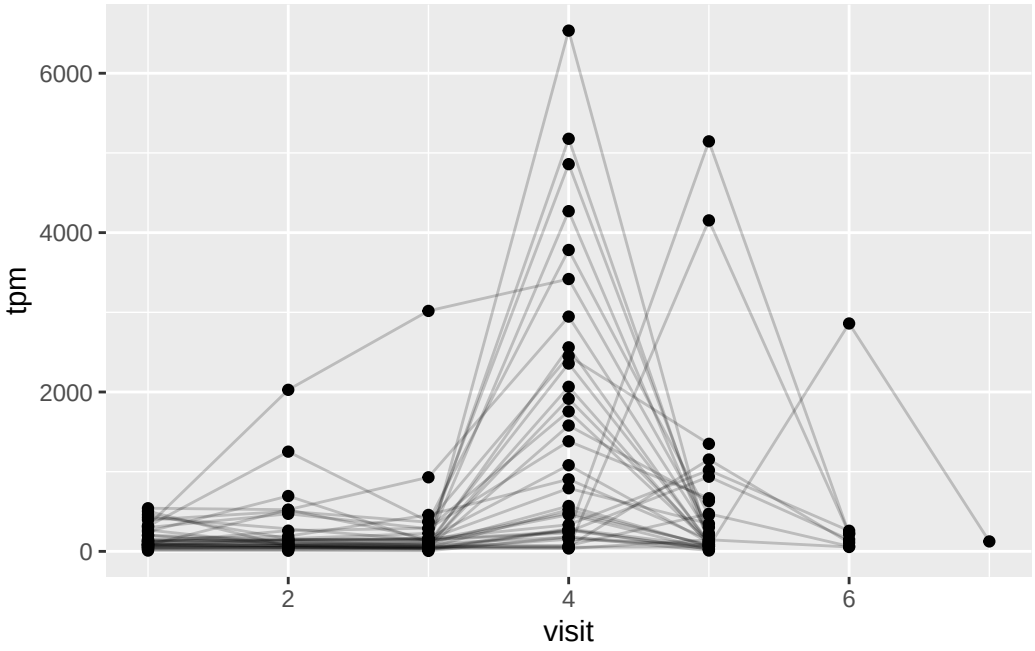
```
rna <- read_json(url, simplifyVector = TRUE)
```

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

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

Q19. Make a plot of the time course of gene expression for IGHG1 gene (i.e. a plot of visit vs. tpm).

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



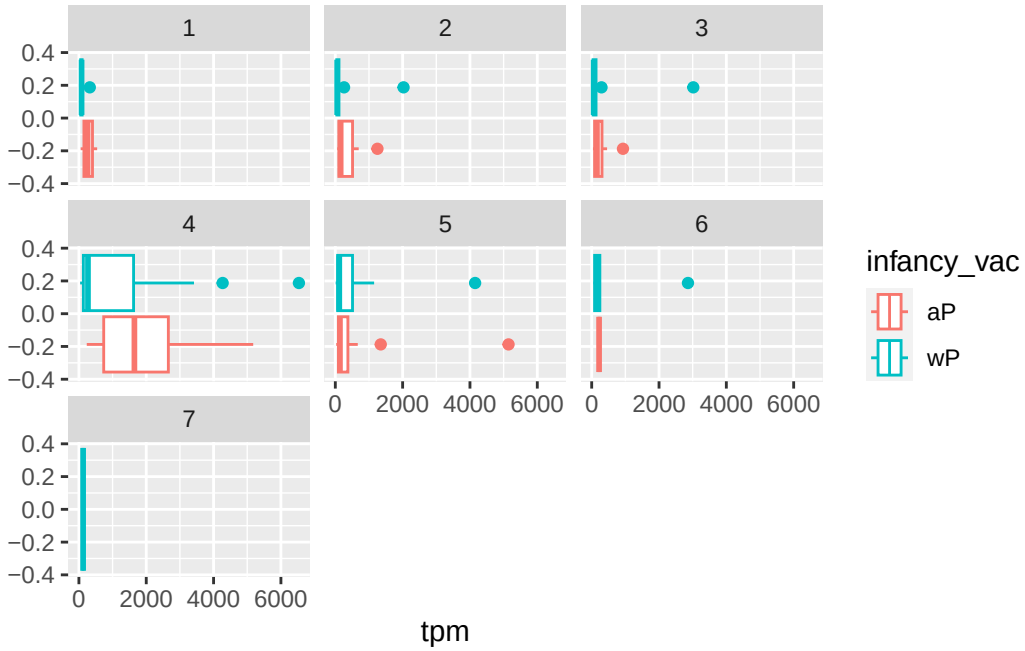
Q20.: What do you notice about the expression of this gene (i.e. when is it at it's maximum level)?

It seems to have a very clear peak at visit #4.

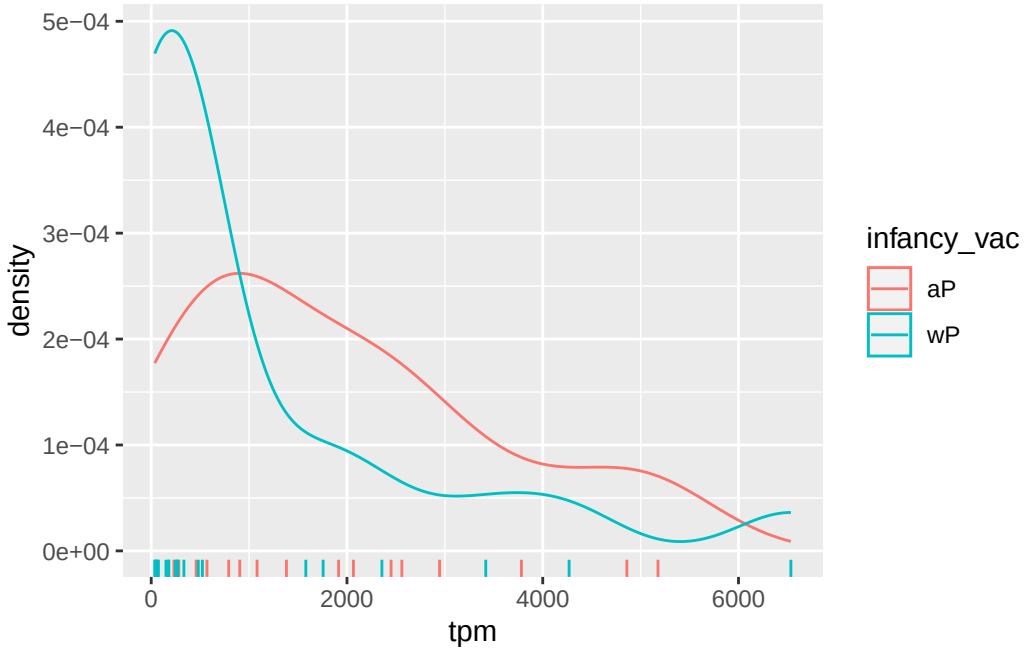
Q21. Does this pattern in time match the trend of antibody titer data? If not, why not?

This adds up with the previous titer data. Levels of antibodies, like PT, started to rise at visit 4. They differ in the levels following week 4. They peak at week 5 in the titer data. This makes sense however, because the antibodies stick around longer than the rna.

```
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()
```



#6. Working with larger datasets [OPTIONAL]

```
# Change for your downloaded file path
rnaseq <- read.csv("~/Downloads/2020LD_rnaseq.csv")

head(rnaseq,3)
```

```
  versioned_ensembl_gene_id specimen_id raw_count tpm
1         ENSG00000229704.1         209         0   0
2         ENSG00000229707.1         209         0   0
3         ENSG00000229708.1         209         0   0
```

```
dim(rnaseq)
```

```
[1] 10502460        4
```

```
n_genes <- table(rnaseq$specimen_id)
head( n_genes , 10)
```

```
    1     3     4     5     6    19    20    21    22    23
58347 58347 58347 58347 58347 58347 58347 58347 58347 58347
```

```
  length(n_genes)
```

```
[1] 180
```

```
  all(n_genes[1]==n_genes)
```

```
[1] TRUE
```

```
  library(tidyr)

  rna_wide <- rnaseq %>%
    select(versioned_ensembl_gene_id, specimen_id, tpm) %>%
    pivot_wider(names_from = specimen_id, values_from=tpm)

  dim(rna_wide)
```

```
[1] 58347   181
```

```
  head(rna_wide[,1:7], 3)
```

```
# A tibble: 3 x 7
  versioned_ensembl_gene_id `209`  `74` `160`  `81` `102` `163`
  <chr>                   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 ENSG00000229704.1           0     0     0     0     0     0
2 ENSG00000229707.1           0     0     0     0     0     0
3 ENSG00000229708.1           0     0     0     0     0     0
```

```
  sessionInfo()
```

```
R version 4.3.2 (2023-10-31)
Platform: aarch64-apple-darwin20 (64-bit)
Running under: macOS Sonoma 14.1.1

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRlapack.dylib;  L

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

time zone: America/Los_Angeles
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] jsonlite_1.8.7  lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1
 [5] dplyr_1.1.4     purrr_1.0.2     readr_2.1.4     tidyr_1.3.0
 [9] tibble_3.2.1    ggplot2_3.4.4   tidyverse_2.0.0 datapasta_3.1.0

loaded via a namespace (and not attached):
 [1] gtable_0.3.4      compiler_4.3.2    tidyselect_1.2.0  scales_1.3.0
 [5] yaml_2.3.7        fastmap_1.1.1     R6_2.5.1          labeling_0.4.3
 [9] generics_0.1.3    knitr_1.45        munsell_0.5.0     pillar_1.9.0
[13] tzdb_0.4.0        rlang_1.1.2       utf8_1.2.4        stringi_1.8.2
[17] xfun_0.41         timechange_0.2.0  cli_3.6.1         withr_2.5.2
[21] magrittr_2.0.3    digest_0.6.33     grid_4.3.2        rstudioapi_0.15.0
[25] hms_1.1.3         lifecycle_1.0.4   vctrs_0.6.4       evaluate_0.23
[29] glue_1.6.2        farver_2.1.1      fansi_1.0.5       colorspace_2.1-0
[33] rmarkdown_2.25    tools_4.3.2       pkgconfig_2.0.3   htmltools_0.5.7
```