Pertussis Resurgence

Vera Sophia Beliaev

3/8/2022

Investigating Pertussis Cases by Year

Web scraping from the CDC website with the help of the datapasta package: let's you copy and paste in data from a website and it will be interpretted as a data frame

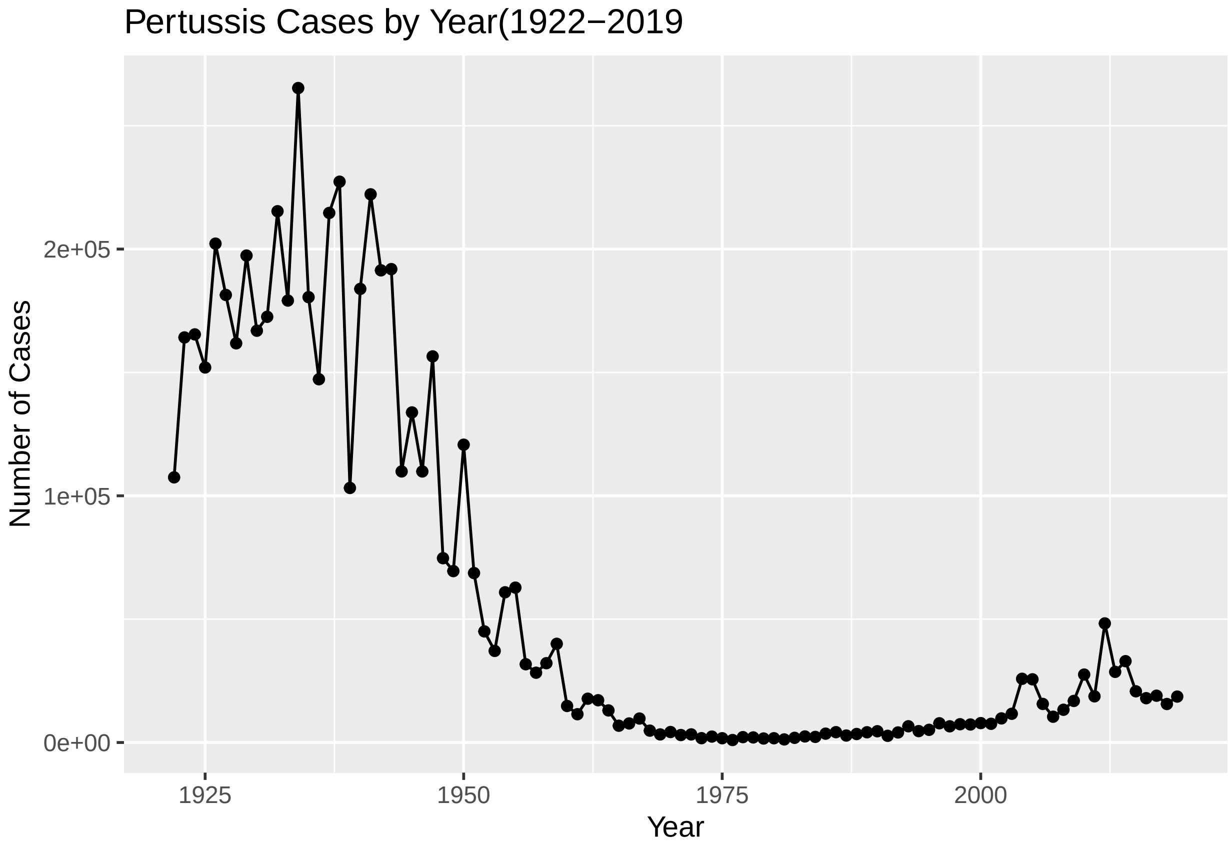
> Q1. With the help of the R "addin" package datapasta assign the CDC pertussis case number data to a data frame called cdc and use ggplot to make a plot of cases numbers over time.

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

```
#put data into clipboard,go to Addins then click paste as dataframe

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),
         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)
       )
```

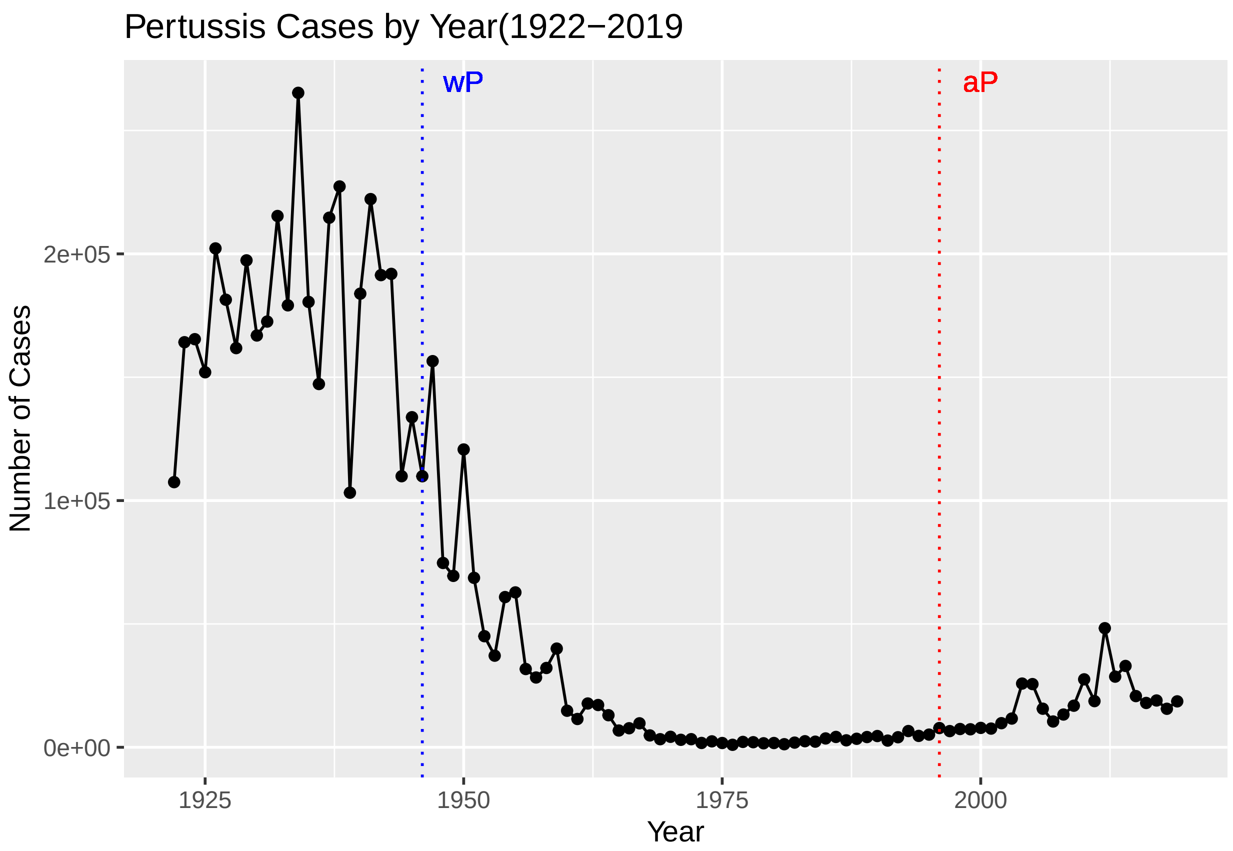
```
ggplot(cdc) + aes(Year, No..Reported.Pertussis.Cases) + geom_point() + geom_line() + labs(x="Year", y=
```



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() + labs(x="Year", y=
```



Q3. Describe what happened after the introduction of the aP vaccine? Do you have a possible explanation for the observed trend? Pertussis cases were increasing after the introduction of the aP vaccine. There could be many explanations including lower rates of vaccination, evolving Bordetella pertussi, waning immunity, more PCR testing, or less efficacy of the aP vaccine compared to the wP vaccine.

Exploring SMI-PB Data

CMI-PB tracks "long-term humoral and cellular immune response data for a large number of individuals who received either DTwP or DTaP combination vaccines in infancy followed by Tdap booster vaccinations"

CMI-PB API gives JSON data which is formatted as a series of key-value pairs (keys/particular words are associated w/ a aprticular value)

JSON format example: { "isotype" : "IgG", "antigen" : "PT" }

to read JSON files, use read_json() fn in jsonlite package or rjson package

```
#jsonlite can simplify JSON key-value pair arrays into R data frames
library(jsonlite)
```

```
#Read the main subject database table, metadata about study participants
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   study_name
## 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 may aP and wP infancy vaccinated subjects are in the dataset? There are 47 aP infancy vaccinated subjects and 49 wP infancy vaccinated subjects in the dataset.

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

```
##
## aP wP
## 47 49
```

> Q5. How many Male and Female subjects/patients are in the dataset? There are 66 female subjects and 30 male subjects in the dataset.

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

```
##
## Female   Male
##     66     30
```

> Q6. What is the breakdown of race and biological sex (e.g. number of Asian females, White males etc…)? The breakdown of race and biological sex is listed in the table below.

```
table(subject$biological_sex, subject$race)
```

```
##
##          American Indian/Alaska Native Asian Black or African American
##   Female                             0    18                         2
##   Male                               1     9                         0
##
##          More Than One Race Native Hawaiian or Other Pacific Islander
##   Female                  8                                         1
##   Male                    2                                         1
##
##          Unknown or Not Reported White
##   Female                      10    27
##   Male                         4    13
```

Side-Note: Working with dates

```
library(lubridate)
```

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

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
today()
```

```
## [1] "2022-03-08"
```

in lubridate, use ymd() to tell the format of the data’s date, and use time_length(, “years) fn to convert days to years

> 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? The average age of wP individuals is 35 years old and the average age of aP individuals is 24 years old.

Problem on Knit: library(dplyr) wp <- subject %>% filter(infancy_vac == “wP”) round(summary(time_length(ap$age, ”year$
$-subject$age, “years”)))

```
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
```

> Q8. Determine the age of all individuals at time of boost? The average age of individuals at the time of boost in 26 years old.

```
#new col w/ age of individs at time of boost in days
subject$boostage <-ymd(subject$date_of_boost) - ymd(subject$year_of_birth)
```

```
round(summary(time_length(subject$boostage, "years")))
```

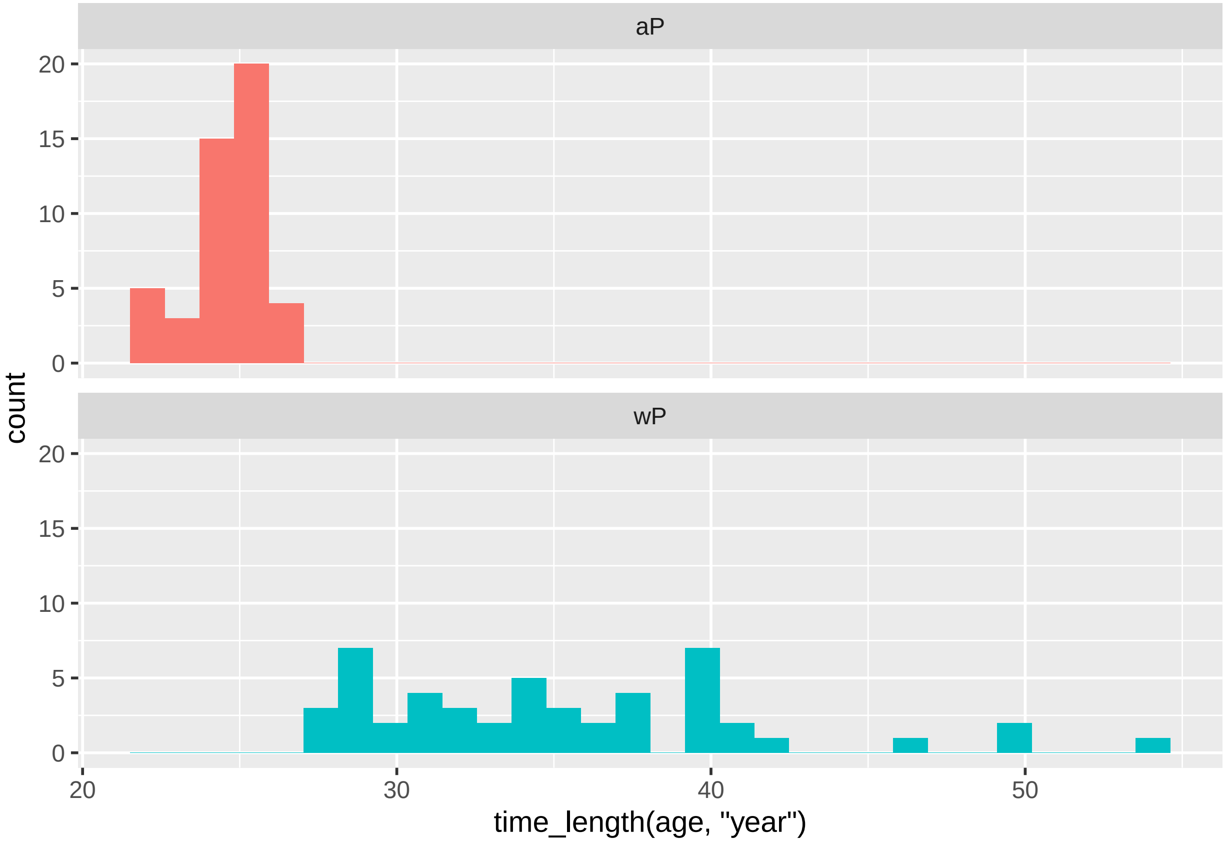
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      19      20      23      26      29      51
```

> Q8. With the help of a faceted boxplot (see below), do you think these two groups are significantly different? Yes, based on the faceted boxplot, the two groups are significantly different.

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

```
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`.
```



Joining Multiple Tables

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

We need to link/join/merge the specimen and subject data frames with dplyr's join() fn

> 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:

```
dim(subject)
```

```
## [1] 96 10
```

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

```
## Joining, by = "subject_id"
```

```
dim(meta)
```

```
## [1] 729  15
```

```
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   study_name
## 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
##     boostage        age
## 1 11212 days 13215 days
## 2 11212 days 13215 days
## 3 11212 days 13215 days
## 4 11212 days 13215 days
## 5 11212 days 13215 days
## 6 11212 days 13215 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.

```
#dim() shows # of rows and cols
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? The number of specimens for each isotype is specified in the table below.

```
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? There are way less number of 8 visit specimens compared to the other visits.

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

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

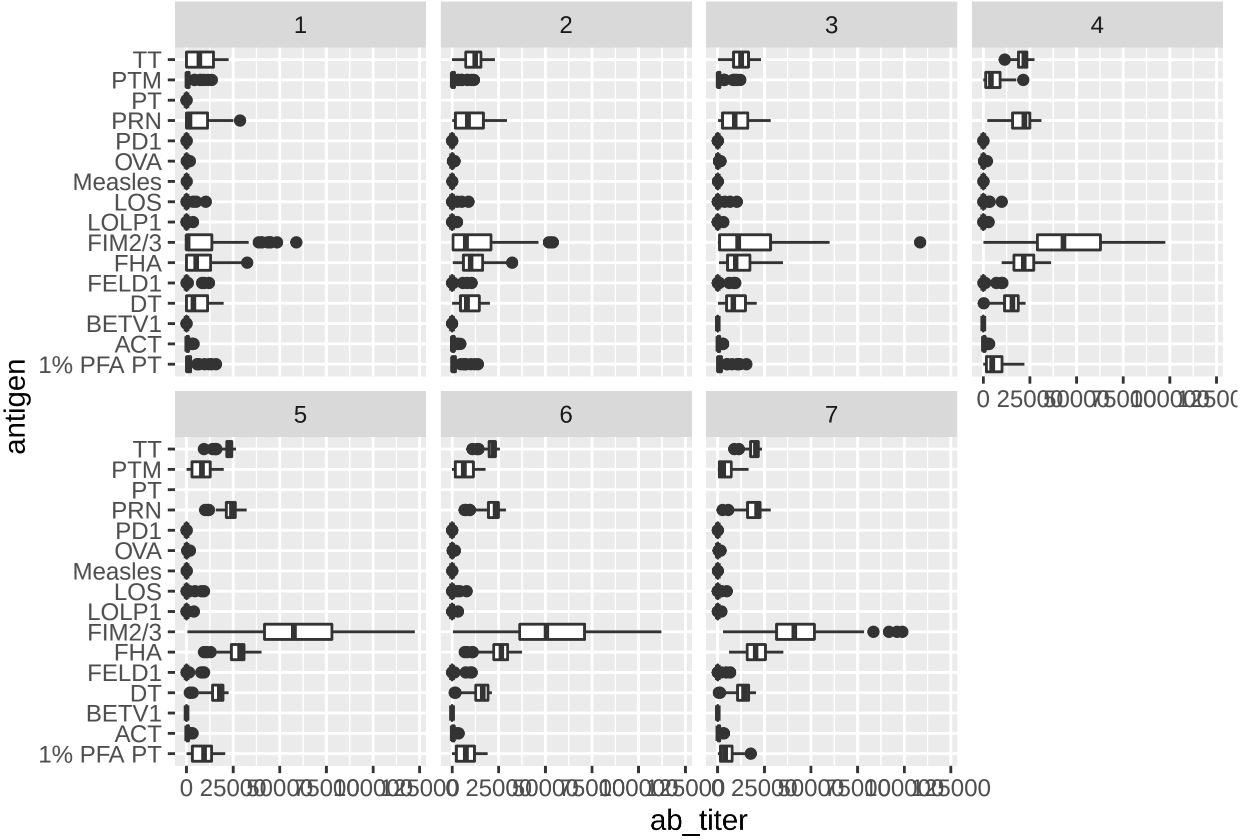
Examine IgG1 Titer Levels

```
#filter for IgG1 isotype, excluding the visit 8 entries
ig1 <- abdata %>% filter(isotype == "IgG1", visit!=8)
head(ig1)
```

```
##   specimen_id isotype is_antigen_specific antigen   ab_titer  unit
## 1           1    IgG1                TRUE     ACT 274.355068 IU/ML
## 2           1    IgG1                TRUE     LOS  10.974026 IU/ML
## 3           1    IgG1                TRUE   FELD1   1.448796 IU/ML
## 4           1    IgG1                TRUE   BETV1   0.100000 IU/ML
## 5           1    IgG1                TRUE   LOLP1   0.100000 IU/ML
## 6           1    IgG1                TRUE Measles  36.277417 IU/ML
##   lower_limit_of_detection subject_id actual_day_relative_to_boost
## 1                 3.848750          1                           -3
## 2                 4.357917          1                           -3
## 3                 2.699944          1                           -3
## 4                 1.734784          1                           -3
## 5                 2.550606          1                           -3
## 6                 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    study_name
## 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
##     boostage        age
## 1 11212 days 13215 days
## 2 11212 days 13215 days
## 3 11212 days 13215 days
## 4 11212 days 13215 days
## 5 11212 days 13215 days
## 6 11212 days 13215 days
```

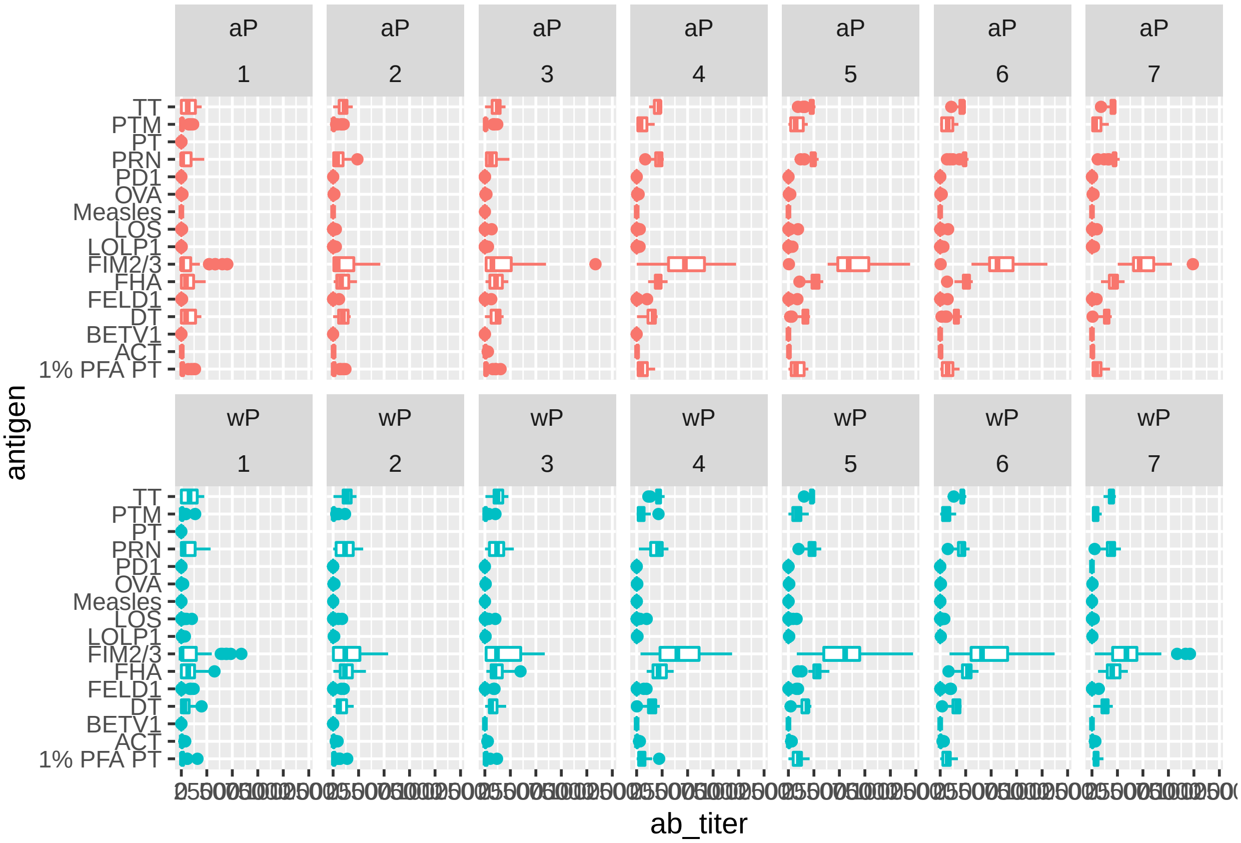
Q13. Complete the following code to make a summary boxplot of Ab titer levels for all antigens:

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



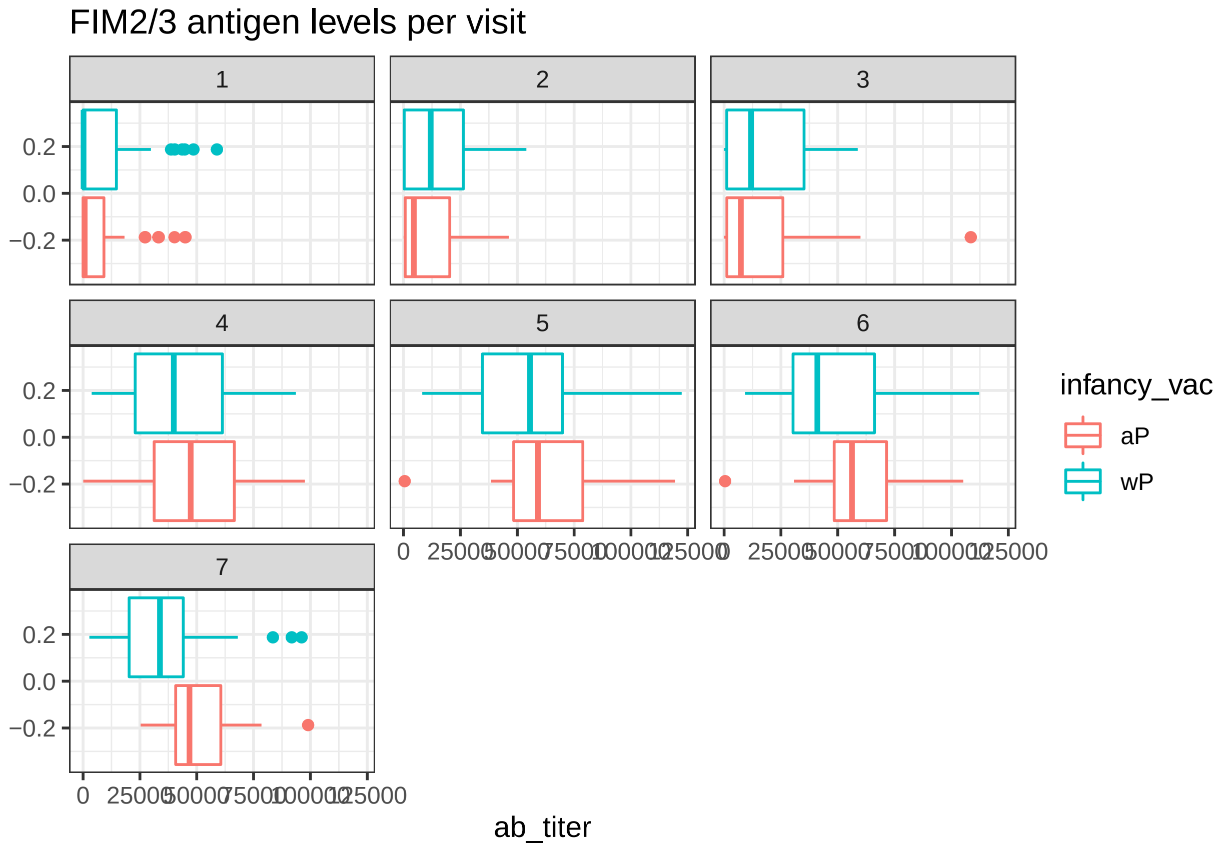
Q14. What antigens show differences in the level of IgG1 antibody titers recognizing them over time? Why these and not others? Antigens such as PT, FIM2/3, and FHA show differences in the level of IgG1 antiody titers recognizing them over time. These may be antigens whose expression changes over time.

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

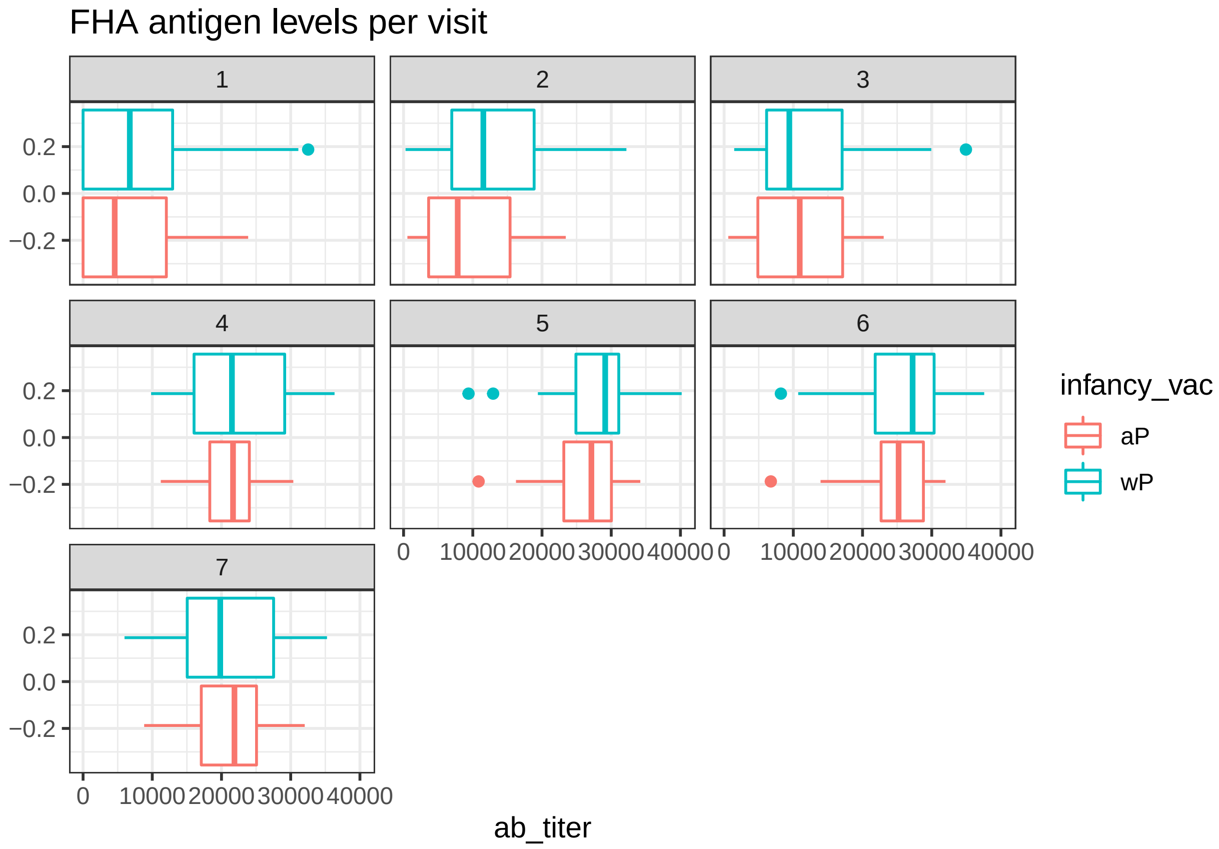


Q15. Filter to pull out only two specific antigens for analysis and create a boxplot for each. You can chose any you like. Below I picked a "control" antigen ("Measles", that is not in our vaccines) and a clear antigen of interest ("FIM2/3", extra-cellular fimbriae proteins from B. pertussis that participate in substrate attachment).

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



```
filter(ig1, antigen == "FHA") %>%
  ggplot() + aes(ab_titer, col=infancy_vac) + geom_boxplot(show.legend = TRUE) + facet_wrap(vars(visit))
```



Q16. What do you notice about these two antigens time course and the FIM2/3 data in particular? The antigens appear to increase over time for the first few visits then decrease over time for the last few visits.

Q17. Do you see any clear difference in aP vs. wP responses? In the case of FIM 2/3 antigen, wP vaccianted appear to have slightly higher antigen levels compared to aP vaccinated. In the case of FHA antigen, aP vaccinated appear to have less variance in antigen levels than wP vaccianted.

Obtaining CMI-PB RNASeq Data

We will obtain RNA-Seq results for specific ENSEMBLE gene identifiers which can be combined with the & character

```
url <- "https://www.cmi-pb.org/api/v2/rnaseq?versioned_ensembl_gene_id=eq.ENSG00000211896.7"

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

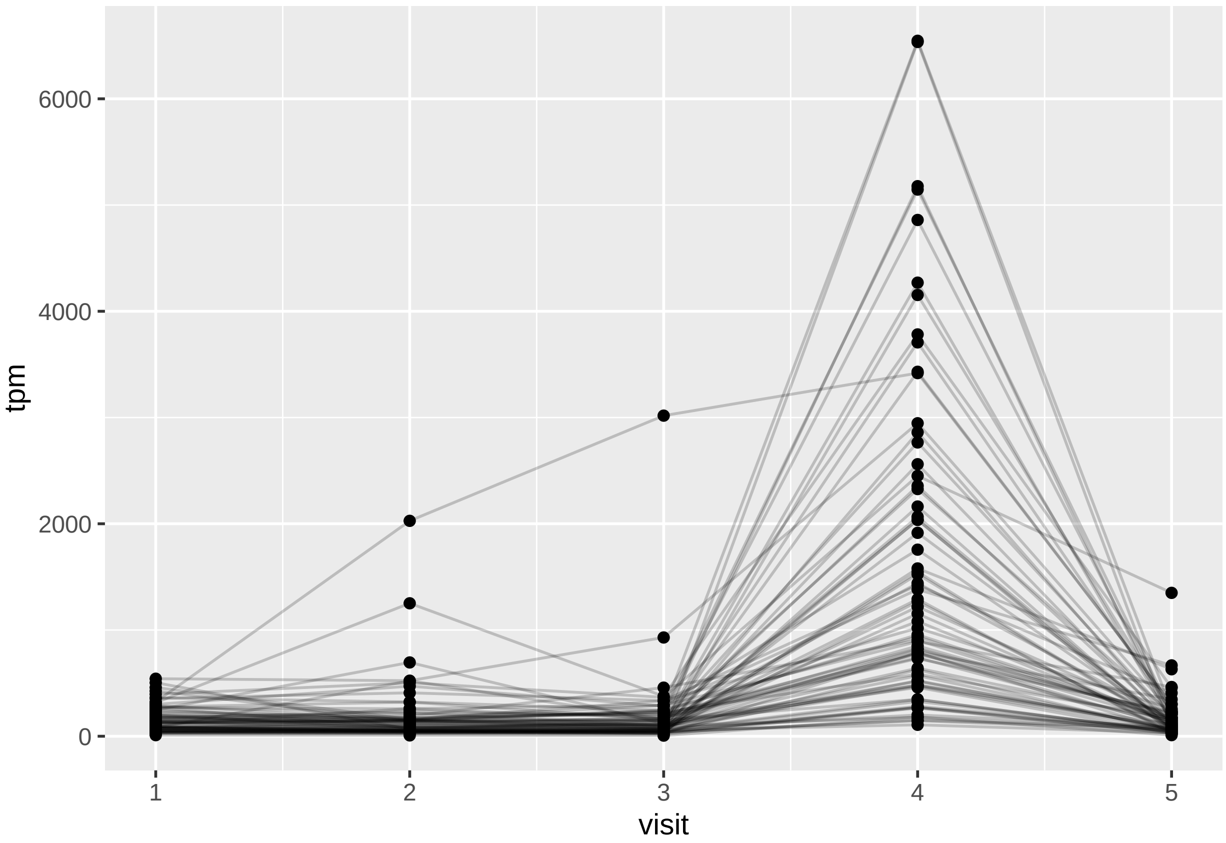
"Join" the rna expression data to our metadata named meta

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

```
## Joining, by = "specimen_id"
```

Q18. Make a plot of the time course of gene expression for IGHG1 gene (i.e. a plot of visit vs. tpm).

```
# Note: tpm is TPM expression values
ggplot(ssrna) + aes(visit, tpm, group=subject_id) + geom_point() + geom_line(alpha=0.2)
```

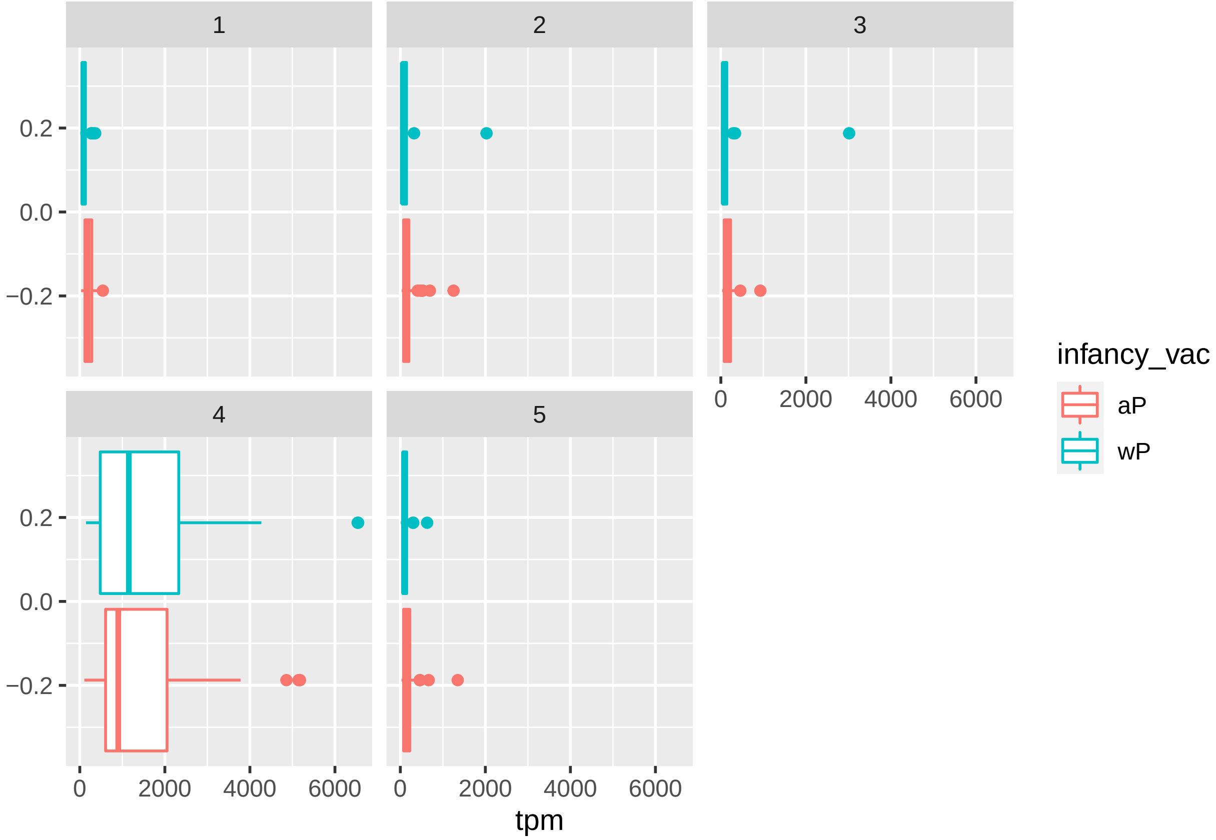


> Q19. What do you notice about the expression of this gene (i.e. when is it at it's maximum level)? The expression of the IGHG1 gene increases then decreases back down over time. It is at its maximum expression when measured at the fourth visit.

Q20. Does this pattern in time match the trend of antibody titer data? If not, why not? No because previously it appeared that antigens reach their peak levels when measured at the fifth visit. This may be because antibodies persist longer in the body than the expression of an individual gene.

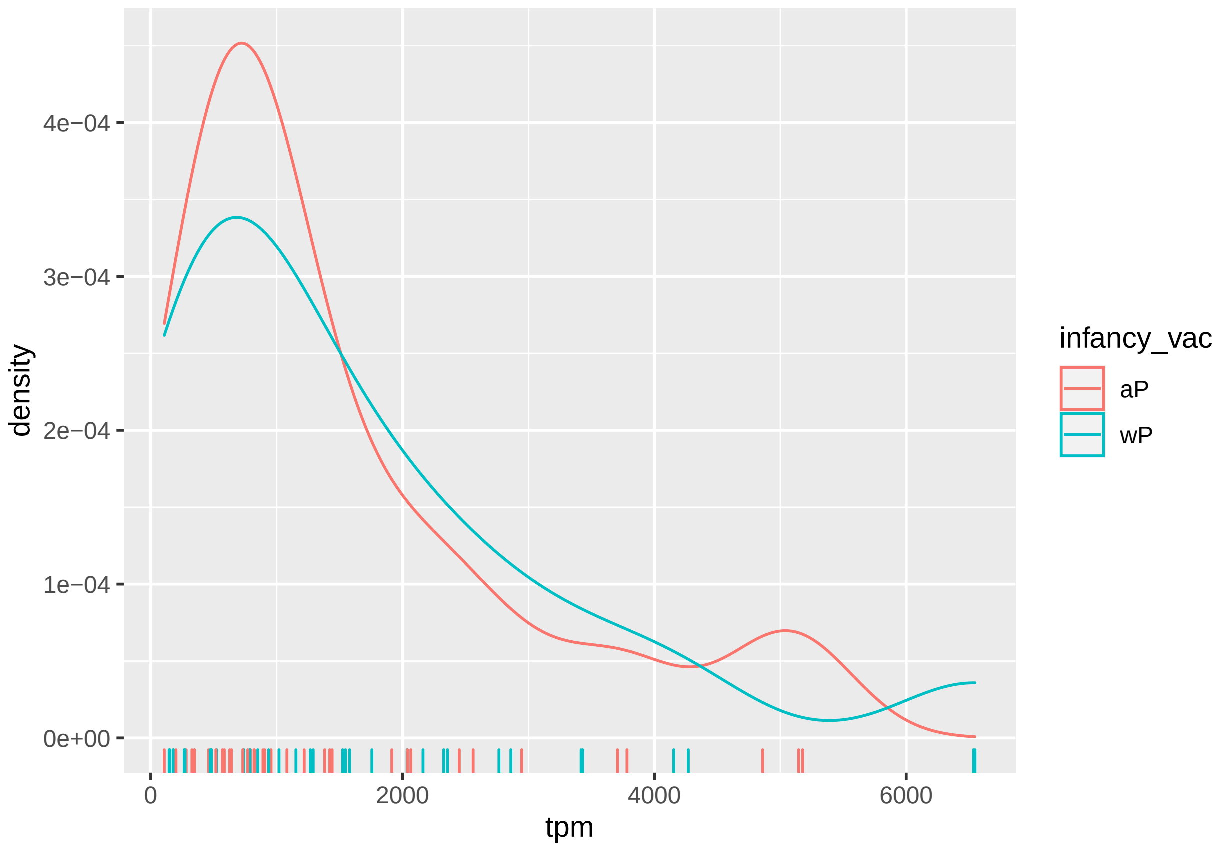
No obvious aP vs wP difference in expression of IGHG1 gene (shown below).

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



Below, we focus on a single visit to evaluate whether there is wP vs aP difference in IGHG1 expression however there is no obvious difference.

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



Working with Larger Datasets

We will use the "2020 longitudinal RNA-Seq data" from CMI-PB, provided as a CSV file

```
# Change for your downloaded file path
rnaseq <- read.csv("/Users/belia/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
```

```
# number of rows and cols in this huge dataset
dim(rnaseq)
```

```
## [1] 10502460        4
```

Working with Long Format Data

The rnadata is in "long format" rather than the conventional wide format In wide format, rows=genes, cols= counts in different experiments

```
#Check how many genes are reported for e/ specimen_id
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
```

```
# Check the number of specimens
length(n_genes)
```

```
## [1] 180
```

```
#Check that the # of genes is the same for all speciments
all(n_genes[1]==n_genes)
```

```
## [1] TRUE
```

Convert to "Wide" Format

use pivot_wider() fn from tidyr package

```
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
```