r— title: “Pertussis Mini Project” author: “Rachel Diao” date: “3/9/2022” output: html_document — Use datapasta to copy and paste data from CDC website link, converts it to a dataframe

```
library(datapasta)
library(ggplot2)
library(jsonlite)
library(lubridate)
```

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

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

```
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
```

```
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,
```

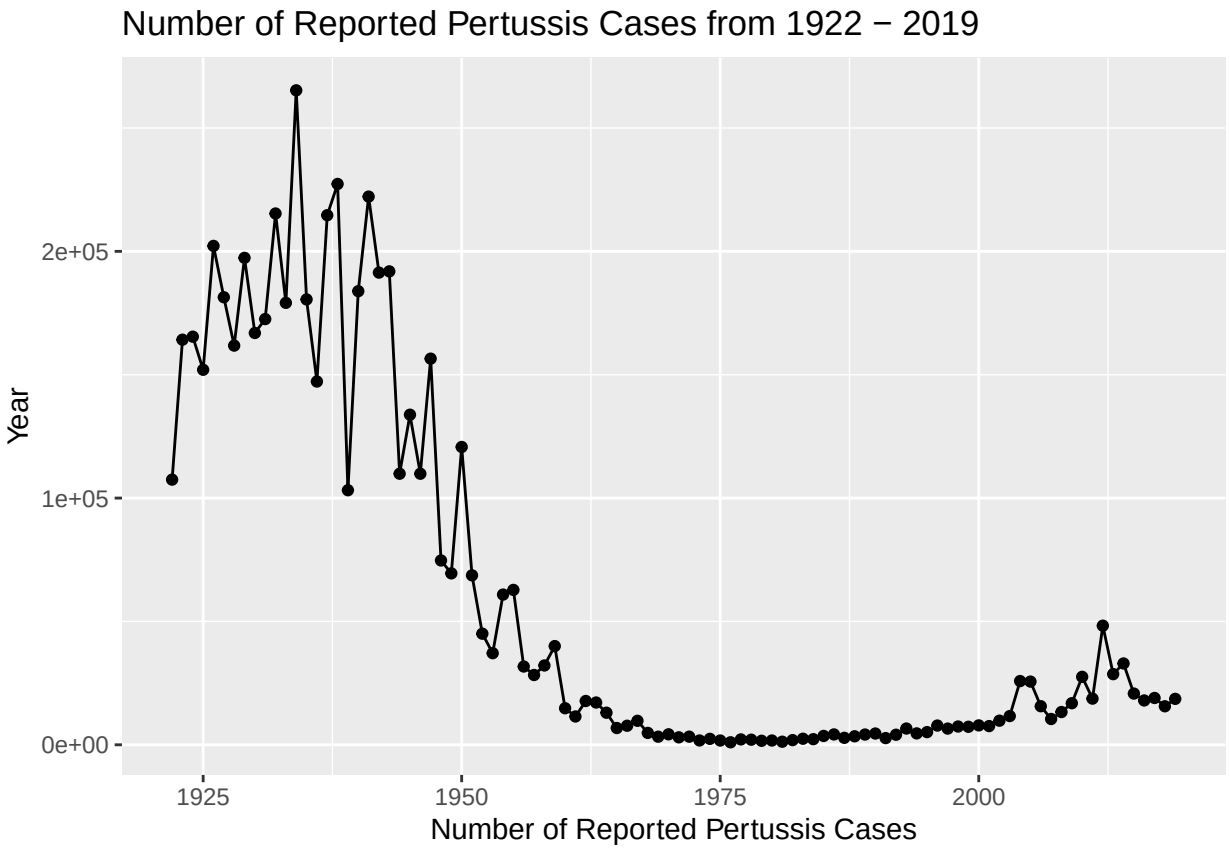
```
                                        180518,147237,214652,227319,103188,
                                        183866,222202,191383,191890,109873,
                                        133792,109860,156517,74715,69479,
                                        120718,68687,45030,37129,60886,
                                        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)
      )
```

Question 1

Plot this dataframe of case numbers over time in ggplot.

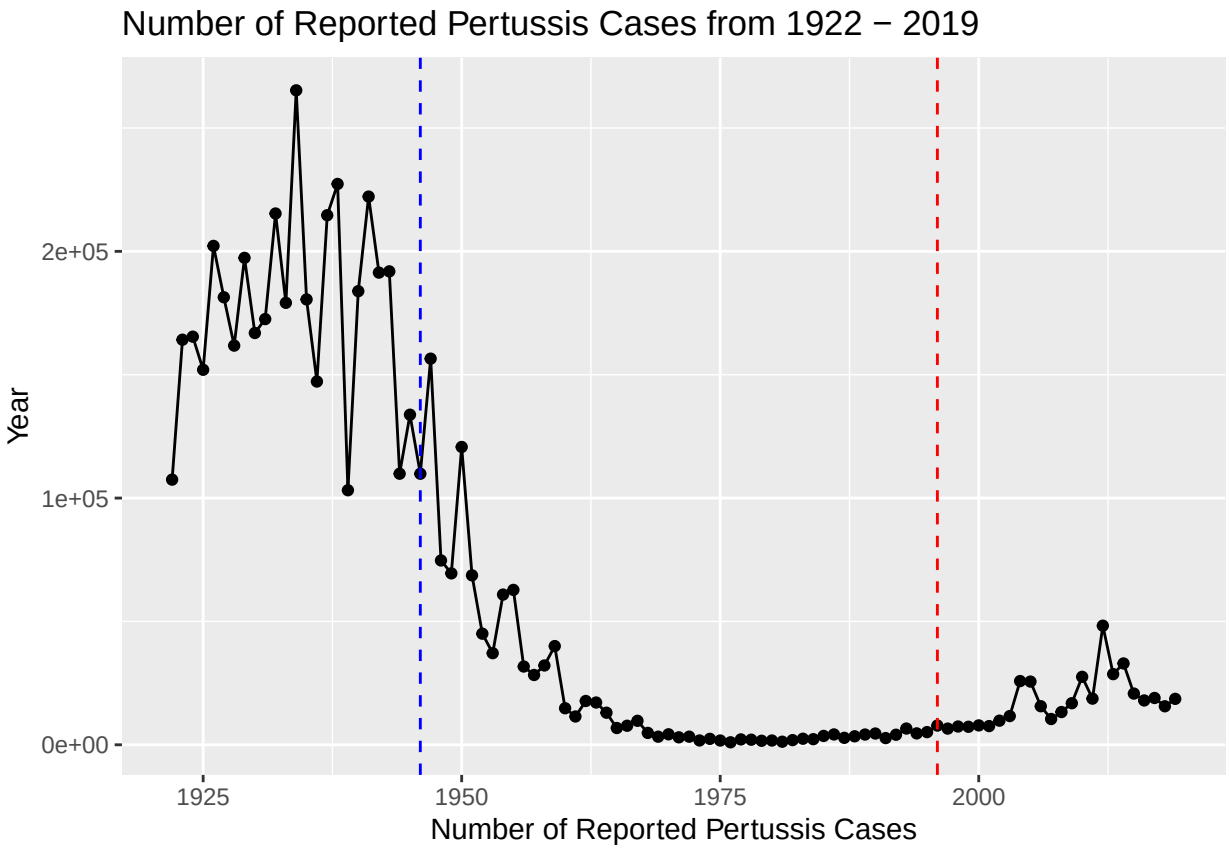
```
pertus <- ggplot(cdc, aes(x=Year, y=No..Reported.Pertussis.Cases)) +
  geom_point() + geom_line() +
  xlab('Number of Reported Pertussis Cases') +
  ylab('Year') +
  ggtitle('Number of Reported Pertussis Cases from 1922 - 2019')

pertus
```



Question 2 Add lines: at 1946 to indicate the introduction of the wP (whole-cell Pertussis) vaccine and the aP (acellular Pertussis) vaccine in 1996.

```
pertus + geom_vline(xintercept=1946, linetype='dashed', col='blue') +
  geom_vline(xintercept=1996, linetype='dashed', col='red')
```



Question 3 After the introduction of the aP version of hte vaccine, very few cases of pertussis remain, but there is slight rise in Pertussis cases starting in the early 2000s. This may be due to the rise of the anti-vaxxer movement in the late 1990s/early 2000s. There also could have been the takeover of a new dominant Pertussis strain, or increases in travel to places with greater Pertussis presence (i.e. countries where vaccination isn't as prominent).

Apparently many teenagers got pertussis in the early 2000s; this may have been because the aP vaccine does not work as well as the old one did, so the immunity has worn off over time (~10 years) in the first cohorts that got vaccinated... but this is very hard to test because it's such a long-term study.

Exploring CMI-PB data

CMI-PB project - will provide information on immune responses over time to wP vs. aP vaccinated individuals.

Need to use jsonlite package to import json type data formats (looks similar to a Python dictionary)

```
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
```

Question 4

49 wP subjects, 47 aP subjects

Question 5

30 male subjects, 66 female subjects

Question 6

American Indian/Alaska Native: 0 females, 1 male Asian: 18 females, 1 male Black or African American: 2 females, 0 males More Than One Race: 8 females, 2 males Native Hawaiian or Other Pacific Islander: 1 female, 1 male Unknown or Not Repoted: 10 females, 4 males White: 27 females, 13 males

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

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

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

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

```
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
```

Question 7

We want to know the average age of wP individuals and aP individuals. Average age of wP individuals: 35 years old Average age of aP individuals: 24 years old

To start, we must first convert all date formats into lubridate objects. The columns are year_of_birth and date_of_boost, all in yyyyy-mm-dd format.

```
subject$year_of_birth <- ymd(subject$year_of_birth)
subject$date_of_boost <- ymd(subject$date_of_boost)
```

Then we can take averages of aP and wP individuals specifically.

```
mean(today() - subject$year_of_birth[which(subject$infancy_vac=='aP')])/365.25
```

```
## Time difference of 24.5026 days
```

```
mean(today() - subject$year_of_birth[which(subject$infancy_vac=='wP')])/365.25
```

```
## Time difference of 35.34705 days
```

Question 8

The average age of all individuals at the time of boost: 25 years old

```
mean(subject$date_of_boost - subject$year_of_birth)/365.25
```

```
## Time difference of 25.60763 days
```

We are skipping through the rest of the date stuff

Joining tables

Retrieve specimen and antibody titer JSON data the same way as before.

```
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)
```

Question 9

Join specimen and subject tables with dplyr join function! Use inner_join() specifically because we only want complete data.

We have multiple specimens per subject (despite only 96 subjects, 729 specimens)

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

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

```
dim(meta)
```

```
## [1] 729  13
```

```
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
```

Question 10

Now join titer to this subject/specimen dataframe.

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

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

```
dim(abdata)
```

```
## [1] 32675    19
```

Question 11

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

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

Question 12

There were only 80 visit 8s, relative to the ~4000+ on the previous visits.

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

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

Examine IgG1

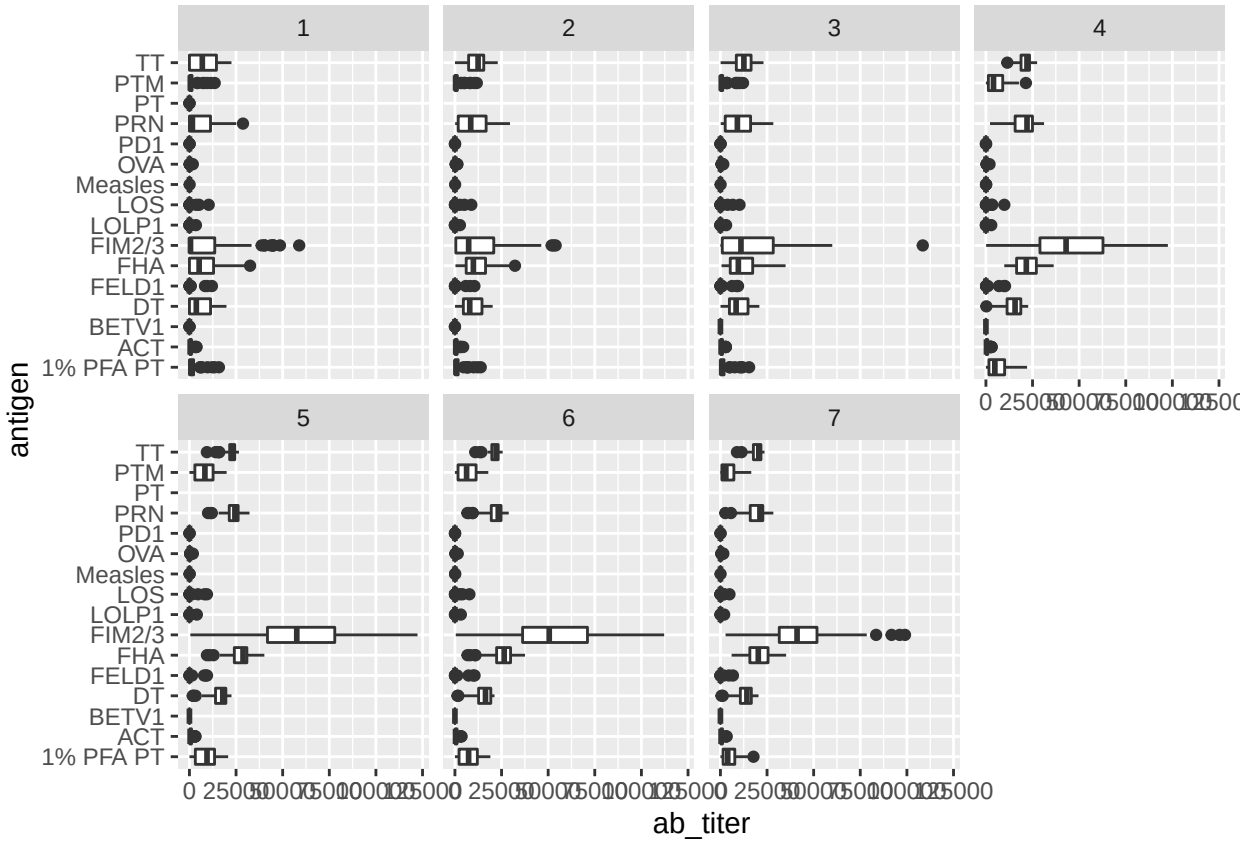
Want to exclude the 8th visit.

```
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
```

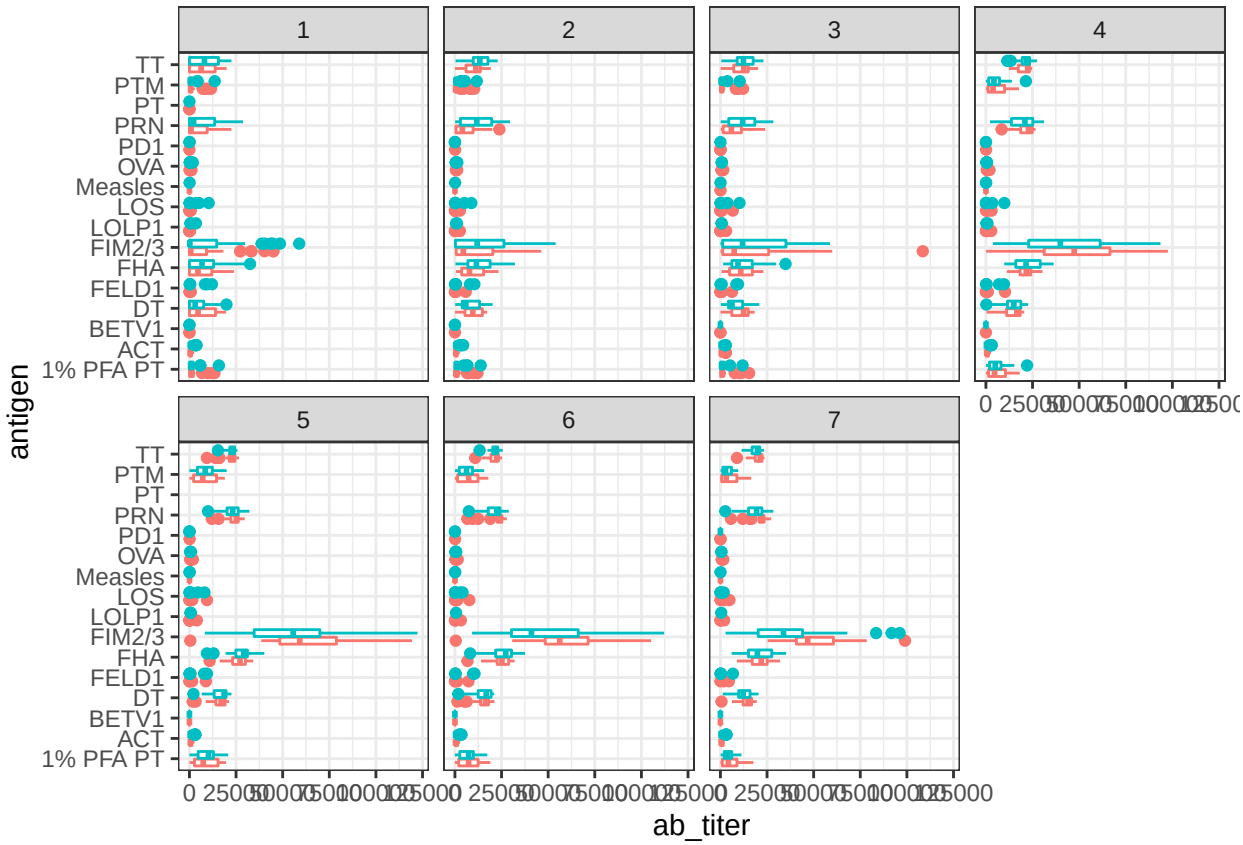
Graph the antibody titer levels for all antigens

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



Now we add in the infancy_vac information (whether they received the wP or aP vaccine) as the color to look at differences in antibody titer between the two.

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


1
2
3
4
5
6
7
antigen
TT
PTM
PT
PRN
PD1
OVA
Measles
LOS
LOLP1
FIM2/3
FHA
FELD1
DT
BETV1
ACT
1% PFA PT
0
25000
50000
75000
100000
125000
ab_titer