

## Class18: Investigating Pertussis Resurgence

Richard Gao (PID: A16490010)

First we will examine and explore Pertussis cas numbers in the US as tracked by the CDC:  
<https://www.cdc.gov/pertussis/surv-reporting/cases-by-year.html>

We can use the `datapasta` package to scrape this data from the website into R:

```
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,2022L,2023L,2024L,2025L,2026L,2027L,2028L,2029L,2030L,2031L,2032L,2033L,2034L,2035L,2036L,2037L,2038L,2039L,2040L,2041L,2042L,2043L,2044L,2045L,2046L,2047L,2048L,2049L,2050L,2051L,2052L,2053L,2054L,2055L,2056L,2057L,2058L,2059L,2060L,2061L,2062L,2063L,2064L,2065L,2066L,2067L,2068L,2069L,2070L,2071L,2072L,2073L,2074L,2075L,2076L,2077L,2078L,2079L,2080L,2081L,2082L,2083L,2084L,2085L,2086L,2087L,2088L,2089L,2090L,2091L,2092L,2093L,2094L,2095L,2096L,2097L,2098L,2099L,2100L,2101L,2102L,2103L,2104L,2105L,2106L,2107L,2108L,2109L,2110L,2111L,2112L,2113L,2114L,2115L,2116L,2117L,2118L,2119L,2120L,2121L,2122L,2123L,2124L,2125L,2126L,2127L,2128L,2129L,2130L,2131L,2132L,2133L,2134L,2135L,2136L,2137L,2138L,2139L,2140L,2141L,2142L,2143L,2144L,2145L,2146L,2147L,2148L,2149L,2150L,2151L,2152L,2153L,2154L,2155L,2156L,2157L,2158L,2159L,2160L,2161L,2162L,2163L,2164L,2165L,2166L,2167L,2168L,2169L,2170L,2171L,2172L,2173L,2174L,2175L,2176L,2177L,2178L,2179L,2180L,2181L,2182L,2183L,2184L,2185L,2186L,2187L,2188L,2189L,2190L,2191L,2192L,2193L,2194L,2195L,2196L,2197L,2198L,2199L,2200L,2201L,2202L,2203L,2204L,2205L,2206L,2207L,2208L,2209L,2210L,2211L,2212L,2213L,2214L,2215L,2216L,2217L,2218L,2219L,2220L,2221L,2222L,2223L,2224L,2225L,2226L,2227L,2228L,2229L,2230L,2231L,2232L,2233L,2234L,2235L,2236L,2237L,2238L,2239L,2240L,2241L,2242L,2243L,2244L,2245L,2246L,2247L,2248L,2249L,2250L,2251L,2252L,2253L,2254L,2255L,2256L,2257L,2258L,2259L,2260L,2261L,2262L,2263L,2264L,2265L,2266L,2267L,2268L,2269L,2270L,2271L,2272L,2273L,2274L,2275L,2276L,2277L,2278L,2279L,2280L,2281L,2282L,2283L,2284L,2285L,2286L,2287L,2288L,2289L,2290L,2291L,2292L,2293L,2294L,2295L,2296L,2297L,2298L,2299L,2300L,2301L,2302L,2303L,2304L,2305L,2306L,2307L,2308L,2309L,2310L,2311L,2312L,2313L,2314L,2315L,2316L,2317L,2318L,2319L,2320L,2321L,2322L,2323L,2324L,2325L,2326L,2327L,2328L,2329L,2330L,2331L,2332L,2333L,2334L,2335L,2336L,2337L,2338L,2339L,2340L,2341L,2342L,2343L,2344L,2345L,2346L,2347L,2348L,2349L,2350L,2351L,2352L,2353L,2354L,2355L,2356L,2357L,2358L,2359L,2360L,2361L,2362L,2363L,2364L,2365L,2366L,2367L,2368L,2369L,2370L,2371L,2372L,2373L,2374L,2375L,2376L,2377L,2378L,2379L,2380L,2381L,2382L,2383L,2384L,2385L,2386L,2387L,2388L,2389L,2390L,2391L,2392L,2393L,2394L,2395L,2396L,2397L,2398L,2399L,2400L,2401L,2402L,2403L,2404L,2405L,2406L,2407L,2408L,2409L,2410L,2411L,2412L,2413L,2414L,2415L,2416L,2417L,2418L,2419L,2420L,2421L,2422L,2423L,2424L,2425L,2426L,2427L,2428L,2429L,2430L,2431L,2432L,2433L,2434L,2435L,2436L,2437L,2438L,2439L,2440L,2441L,2442L,2443L,2444L,2445L,2446L,2447L,2448L,2449L,2450L,2451L,2452L,2453L,2454L,2455L,2456L,2457L,2458L,2459L,2460L,2461L,2462L,2463L,2464L,2465L,2466L,2467L,2468L,2469L,2470L,2471L,2472L,2473L,2474L,2475L,2476L,2477L,2478L,2479L,2480L,2481L,2482L,2483L,2484L,2485L,2486L,2487L,2488L,2489L,2490L,2491L,2492L,2493L,2494L,2495L,2496L,2497L,2498L,2499L,2500L,2501L,2502L,2503L,2504L,2505L,2506L,2507L,2508L,2509L,2510L,2511L,2512L,2513L,2514L,2515L,2516L,2517L,2518L,2519L,2520L,2521L,2522L,2523L,2524L,2525L,2526L,2527L,2528L,2529L,2530L,2531L,2532L,2533L,2534L,2535L,2536L,2537L,2538L,2539L,2540L,2541L,2542L,2543L,2544L,2545L,2546L,2547L,2548L,2549L,2550L,2551L,2552L,2553L,2554L,2555L,2556L,2557L,2558L,2559L,2560L,2561L,2562L,2563L,2564L,2565L,2566L,2567L,2568L,2569L,2570L,2571L,2572L,2573L,2574L,2575L,2576L,2577L,2578L,2579L,2580L,2581L,2582L,2583L,2584L,2585L,2586L,2587L,2588L,2589L,2590L,2591L,2592L,2593L,2594L,2595L,2596L,2597L,2598L,2599L,2600L,2601L,26
```

```
head(cdc)
```

```

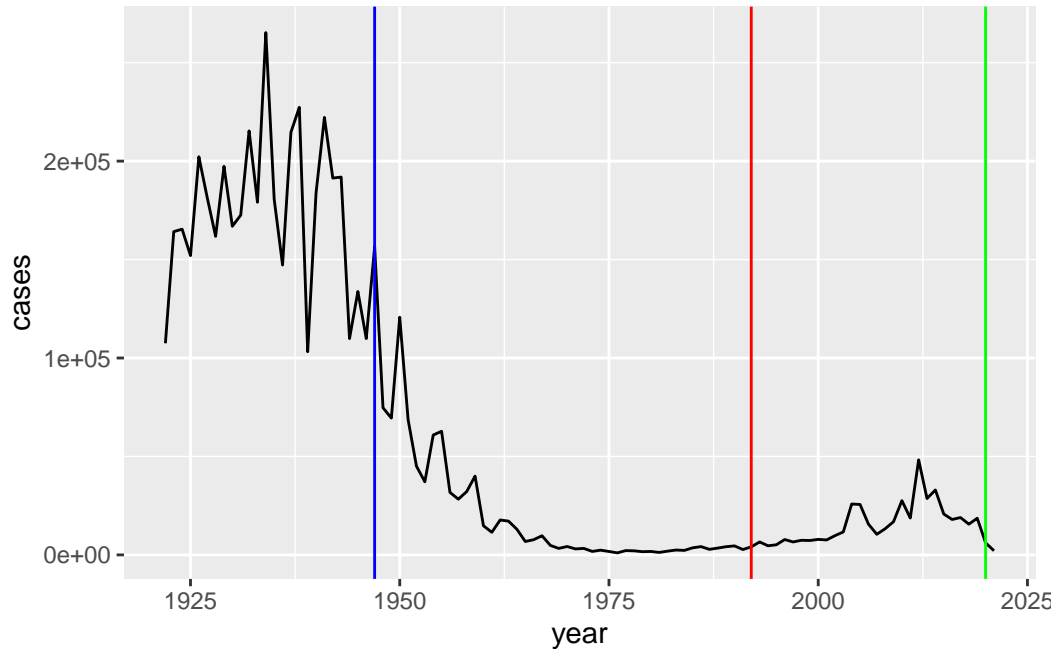
year  cases
1 1922 107473
2 1923 164191
3 1924 165418
4 1925 152003
5 1926 202210
6 1927 181411

```

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(ggplot2)
```

```
ggplot(cdc, aes(year, cases)) +
  geom_line() +
  geom_vline(xintercept=1947, col="blue") +
  geom_vline(xintercept=1992, col="red") +
  geom_vline(xintercept=2020, col="green")
```



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?

I notice that between the 1946 introduction of the wP vaccine and the 1996 switch to the aP vaccine, the number of pertussis cases dropped significantly and stayed low.

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, the number of pertussis cases resurged. A possible explanation for this is that the immunity gained from the aP vaccine doesn't last very long compared to the original wP vaccine.

Access data from the CMI-PB project

This database (like many modern projects) uses an API to return JSON format data.

We will use the R package `jsonlite`.

```
library(jsonlite)
subject <- read_json("http://cmi-pb.org/api/subject",
                     simplifyVector = TRUE)
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

Q4. How many aP and wP infancy vaccinated subjects are in the dataset?

```
sum(subject$infancy_vac == "wP")
```

```
[1] 58
```

```
sum(subject$infancy_vac == "aP")
```

```
[1] 60
```

or just do:

```
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

```
subject$year_of_birth
```

```
[1] "1986-01-01" "1968-01-01" "1983-01-01" "1988-01-01" "1991-01-01"
[6] "1988-01-01" "1981-01-01" "1985-01-01" "1996-01-01" "1982-01-01"
[11] "1986-01-01" "1982-01-01" "1997-01-01" "1993-01-01" "1989-01-01"
[16] "1987-01-01" "1980-01-01" "1997-01-01" "1994-01-01" "1981-01-01"
[21] "1983-01-01" "1985-01-01" "1991-01-01" "1992-01-01" "1988-01-01"
[26] "1983-01-01" "1997-01-01" "1982-01-01" "1997-01-01" "1988-01-01"
[31] "1989-01-01" "1997-01-01" "1990-01-01" "1983-01-01" "1991-01-01"
[36] "1997-01-01" "1998-01-01" "1997-01-01" "1985-01-01" "1994-01-01"
[41] "1985-01-01" "1997-01-01" "1998-01-01" "1998-01-01" "1997-01-01"
[46] "1998-01-01" "1996-01-01" "1998-01-01" "1997-01-01" "1997-01-01"
[51] "1997-01-01" "1998-01-01" "1998-01-01" "1997-01-01" "1997-01-01"
[56] "1997-01-01" "1996-01-01" "1997-01-01" "1997-01-01" "1997-01-01"
[61] "1987-01-01" "1993-01-01" "1995-01-01" "1993-01-01" "1990-01-01"
[66] "1976-01-01" "1972-01-01" "1972-01-01" "1990-01-01" "1998-01-01"
[71] "1998-01-01" "1991-01-01" "1995-01-01" "1995-01-01" "1998-01-01"
[76] "1998-01-01" "1988-01-01" "1993-01-01" "1987-01-01" "1992-01-01"
[81] "1993-01-01" "1998-01-01" "1999-01-01" "1997-01-01" "2000-01-01"
[86] "1998-01-01" "2000-01-01" "2000-01-01" "1997-01-01" "1999-01-01"
[91] "1998-01-01" "2000-01-01" "1996-01-01" "1999-01-01" "1998-01-01"
[96] "2000-01-01" "1986-01-01" "1993-01-01" "1999-01-01" "2001-01-01"
[101] "2003-01-01" "2003-01-01" "1994-01-01" "1989-01-01" "1994-01-01"
[106] "1996-01-01" "1998-01-01" "1995-01-01" "1989-01-01" "1997-01-01"
[111] "1996-01-01" "1996-01-01" "1996-01-01" "1990-01-01" "2002-01-01"
[116] "2000-01-01" "1994-01-01" "1998-01-01"
```

## Side-Note: Working with dates

We can use the lubridate package to ease the pain of doing math with dates.

```
library(lubridate)
```

Attaching package: 'lubridate'

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

```
date, intersect, setdiff, union
```

```
today()
```

```
[1] "2024-03-07"
```

```
today() - ymd("2002-11-19")
```

Time difference of 7779 days

```
time_length(today() - ymd("2002-11-19"), "years")
```

```
[1] 21.29774
```

So what is the age of everyone on our dataset.

```
subject$age <- time_length( today() - ymd(subject$year_of_birth), "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?

```
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
```

```
wPsubject <- filter(subject, infancy_vac == "wP")
head(wPsubject)
```

	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	38.17933
2	1968-01-01	2019-01-28	2020_dataset	56.18070
3	1983-01-01	2016-10-10	2020_dataset	41.18001
4	1988-01-01	2016-08-29	2020_dataset	36.18070
5	1991-01-01	2016-08-29	2020_dataset	33.18001
6	1988-01-01	2016-10-10	2020_dataset	36.18070

```
mean(wPsubject$age)
```

```
[1] 36.57618
```

```
aPsubject <- filter(subject, infancy_vac == "aP")
mean(aPsubject$age)
```

```
[1] 26.27944
```

```
t.test(wPsubject$age, aPsubject$age)
```

### Welch Two Sample t-test

```
data: wPsubject$age and aPsubject$age
t = 12.436, df = 65.411, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 8.643385 11.950080
sample estimates:
mean of x mean of y
36.57618 26.27944
```

Average age of wP individuals is 36.58 yrs old and that of aP individuals is 26.28 yrs old. Yes they are significantly different with a p-value < 2.2e-16.

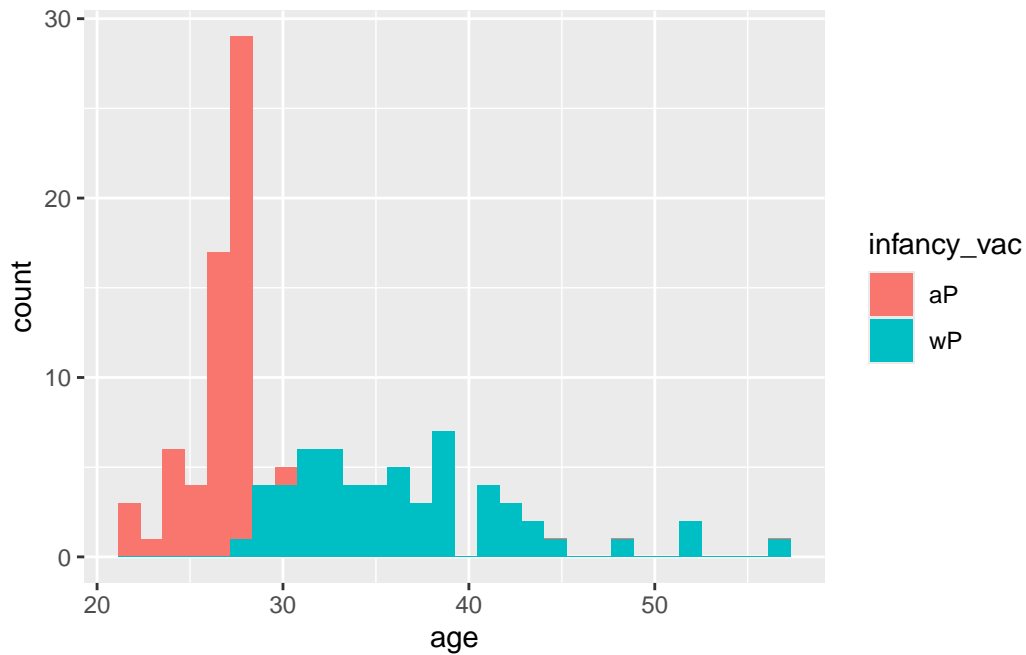
Q8. Determine the age of all individuals at time of boost?

```
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
```

```
ggplot(subject) +
  aes(age, fill = infancy_vac) +
  geom_histogram()
```

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

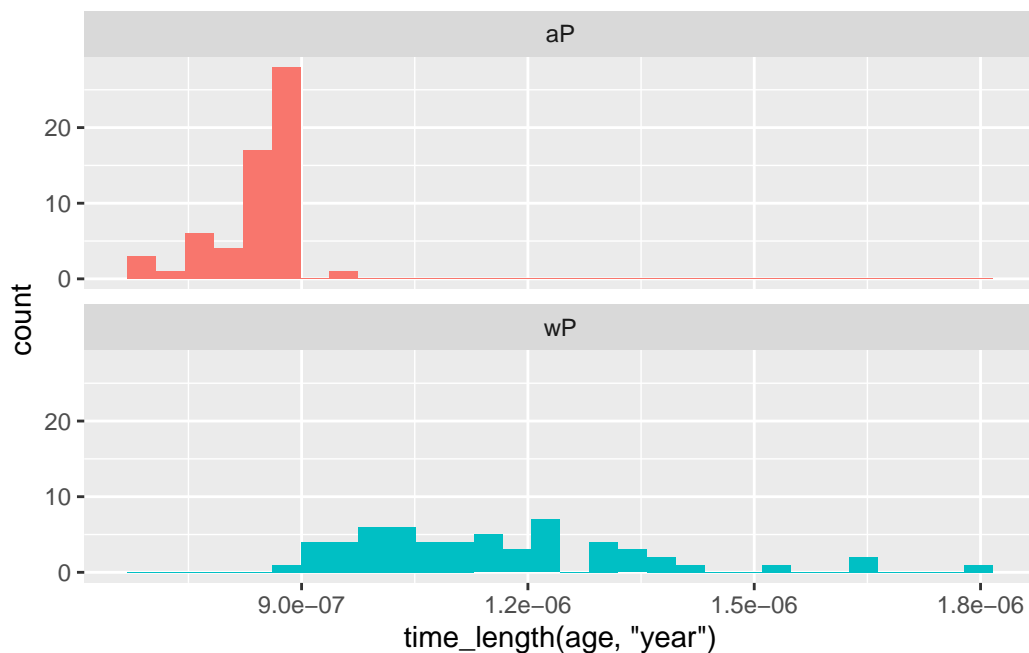


Q9. With the help of a faceted boxplot or histogram (see below), do you think these two groups are significantly different?

```
ggplot(subject) +
  aes(time_length(age, "year"),
      fill=as.factor(infancy_vac)) +
  geom_histogram(show.legend=FALSE) +
  facet_wrap(vars(infancy_vac), nrow=2)
```

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.





Yes, those who received the wP vaccine seem to be significantly older than those who received the aP vaccine.

## Get more data from CMI-PB

```
specimen <- read_json("http://cmi-pb.org/api/specimen", simplifyVector = T)
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

We need to **join** these two tables (subject and specimen) to make a single new “meta” table with all our metadata. We will use the `dplyr` join functions to do this:

Q10. 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:

```
library(dplyr)

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

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

```
head(meta)
```

	subject_id	infancy_vac	biological_sex	ethnicity	race
1	1	wP	Female Not Hispanic or Latino	White	
2	1	wP	Female Not Hispanic or Latino	White	
3	1	wP	Female Not Hispanic or Latino	White	
4	1	wP	Female Not Hispanic or Latino	White	
5	1	wP	Female Not Hispanic or Latino	White	
6	1	wP	Female Not Hispanic or Latino	White	

	year_of_birth	date_of_boost	dataset	age	specimen_id
1	1986-01-01	2016-09-12	2020_dataset	38.17933	1
2	1986-01-01	2016-09-12	2020_dataset	38.17933	2
3	1986-01-01	2016-09-12	2020_dataset	38.17933	3
4	1986-01-01	2016-09-12	2020_dataset	38.17933	4
5	1986-01-01	2016-09-12	2020_dataset	38.17933	5
6	1986-01-01	2016-09-12	2020_dataset	38.17933	6

	actual_day_relative_to_boost	planned_day_relative_to_boost	specimen_type
1	-3	0	Blood
2	1	1	Blood
3	3	3	Blood
4	7	7	Blood
5	11	14	Blood
6	32	30	Blood

	visit
1	1

```
2    2
3    3
4    4
5    5
6    6
```

Now we can read some of the other data from CMI-PB

```
ab_titer <- read_json("http://cmi-pb.org/api/v4/plasma_ab_titer",
                      simplifyVector = T)

head(ab_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

One more `inner_join()` to add all our metadata in `meta` on to our `ab_data` table:

Q11. 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(ab_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	infancy_vac	biological_sex
1	UG/ML	2.096133	1	wP	Female
2	IU/ML	29.170000	1	wP	Female
3	IU/ML	0.530000	1	wP	Female
4	IU/ML	6.205949	1	wP	Female
5	IU/ML	4.679535	1	wP	Female
6	IU/ML	2.816431	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	actual_day_relative_to_boost	planned_day_relative_to_boost
1	38.17933	-3	0
2	38.17933	-3	0
3	38.17933	-3	0
4	38.17933	-3	0
5	38.17933	-3	0
6	38.17933	-3	0

	specimen_type	visit
1	Blood	1
2	Blood	1
3	Blood	1
4	Blood	1
5	Blood	1
6	Blood	1

```
dim(abdata)
```

```
[1] 41775    21
```

Our first exploratory plot:

Q12. How many specimens (i.e. entries in abdata) do we have for each isotype?

```
table(abdata$antigen)
```

ACT	BETV1	DT	FELD1	FHA	FIM2/3	LOLP1	LOS	Measles	OVA
1970	1970	3435	1970	3829	3435	1970	1970	1970	3435
PD1	PRN	PT	PTM	Total	TT				
1970	3829	3829	1970	788	3435				

Q13. 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	2170

We have less rows for more recent years.

## Examine IgG Ab titer levels

Now using our joined/merged/linked abdata dataset filter() for IgG isotype.

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

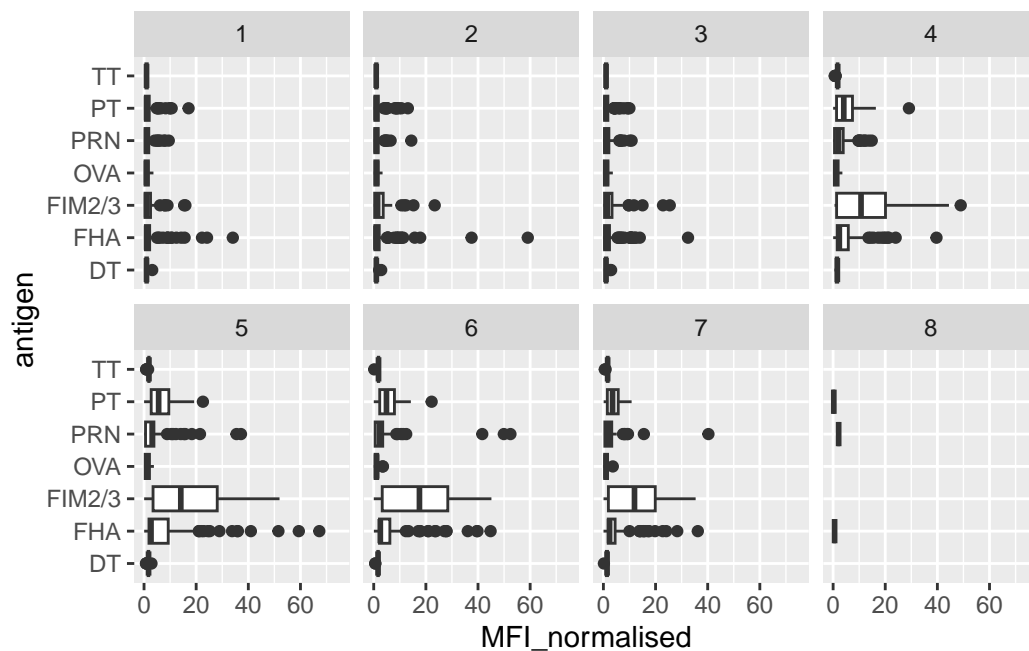
	specimen_id	isotype	is_antigen_specific	antigen	MFI	MFI_normalised
1	1	IgG	TRUE	PT	68.56614	3.736992
2	1	IgG	TRUE	PRN	332.12718	2.602350
3	1	IgG	TRUE	FHA	1887.12263	34.050956
4	19	IgG	TRUE	PT	20.11607	1.096366
5	19	IgG	TRUE	PRN	976.67419	7.652635
6	19	IgG	TRUE	FHA	60.76626	1.096457
	unit	lower_limit_of_detection	subject_id	infancy_vac	biological_sex	
1	IU/ML	0.530000	1	wP	Female	
2	IU/ML	6.205949	1	wP	Female	
3	IU/ML	4.679535	1	wP	Female	

4	IU/ML	0.530000	3	wP	Female
5	IU/ML	6.205949	3	wP	Female
6	IU/ML	4.679535	3	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	actual_day_relative_to_boost	planned_day_relative_to_boost		
1	38.17933		-3		0
2	38.17933		-3		0
3	38.17933		-3		0
4	41.18001		-3		0
5	41.18001		-3		0
6	41.18001		-3		0
	specimen_type	visit			
1	Blood	1			
2	Blood	1			
3	Blood	1			
4	Blood	1			
5	Blood	1			
6	Blood	1			

Q14. 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 outside the scale range (`stat\_boxplot()`).

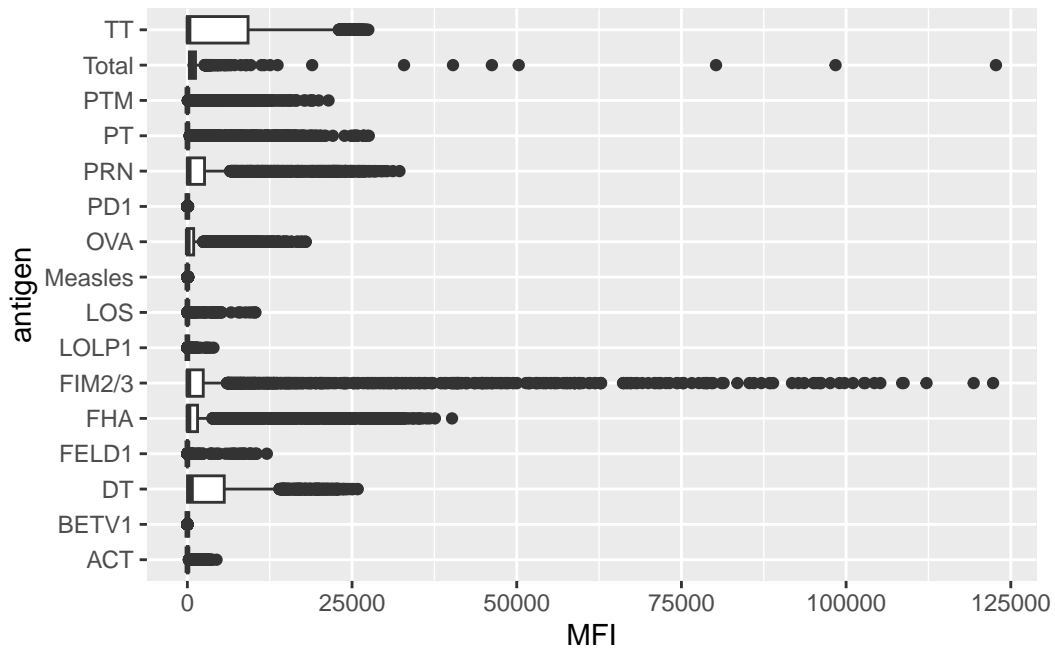


Q15. What antigens show differences in the level of IgG antibody titers recognizing them over time? Why these and not others?

PT, PRN, FIM2/3, and FHA. These were included in the vaccine.

```
ggplot(abdata) +
  aes(MFI, antigen) +
  geom_boxplot()
```

Warning: Removed 1 row containing non-finite outside the scale range (``stat_boxplot()``).



Why are certain antigens and not others very variable in their detected levels here?

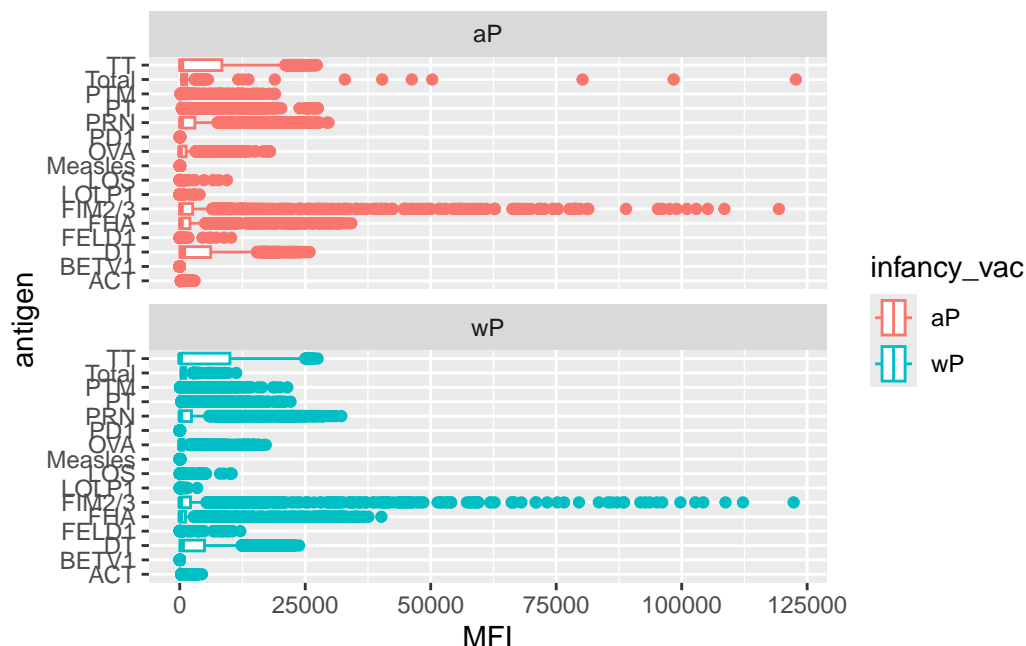
The vaccine only contains certain antigens.

Can you facet or even just color by `infancy_vac`? Is there some difference?

```
ggplot(abdata) +
  aes(MFI, antigen, col=infancy_vac) +
  geom_boxplot() +
  facet_wrap(vars(infancy_vac), nrow=2)
```

Warning: Removed 1 row containing non-finite outside the scale range (``stat_boxplot()``).





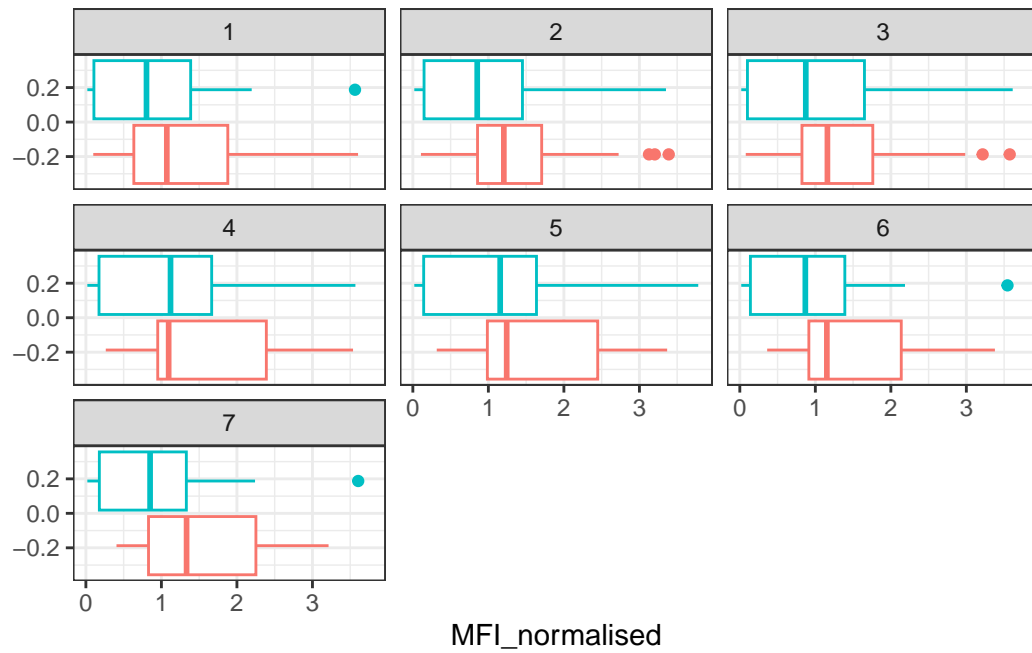
There are potentially some differences here but in general it is hard to tell with this whole dataset overview...

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

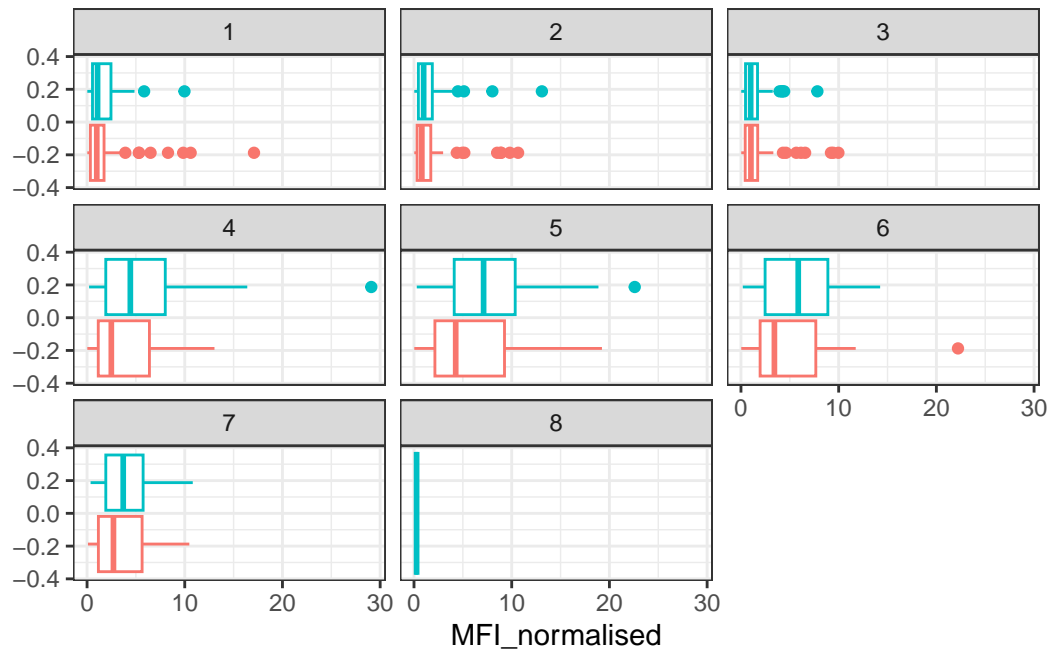
```
2020_dataset 2021_dataset 2022_dataset
      31520         8085         2170
```

Q16. 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*).

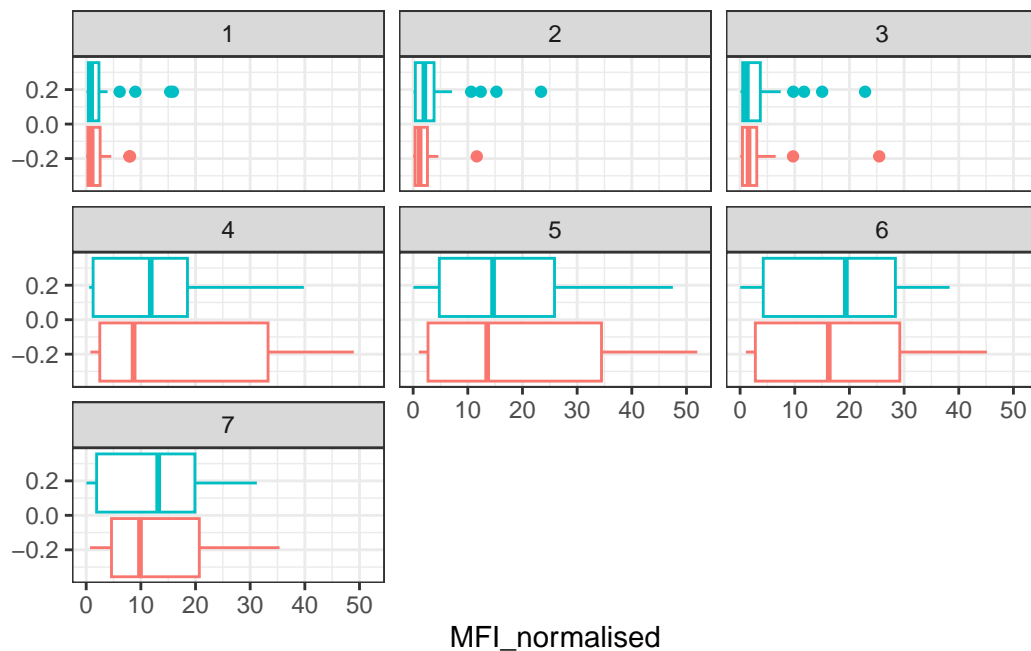
```
filter(igg, antigen=="OVA") %>%
  ggplot() +
  aes(MFI_normalised, col=infancy_vac) +
  geom_boxplot(show.legend = F) +
  facet_wrap(vars(visit)) +
  theme_bw()
```



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



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



Q17. What do you notice about these two antigens time courses and the PT data in particular?

PT increases for a while then has a sharp drop off while FIM 2/3 steadily increases over time and only slightly decreases.

Let's focus in on just the 2021\_dataset.

```
abdata.21 <- filter(abdata, dataset == "2021_dataset")
table(abdata.21$dataset)
```

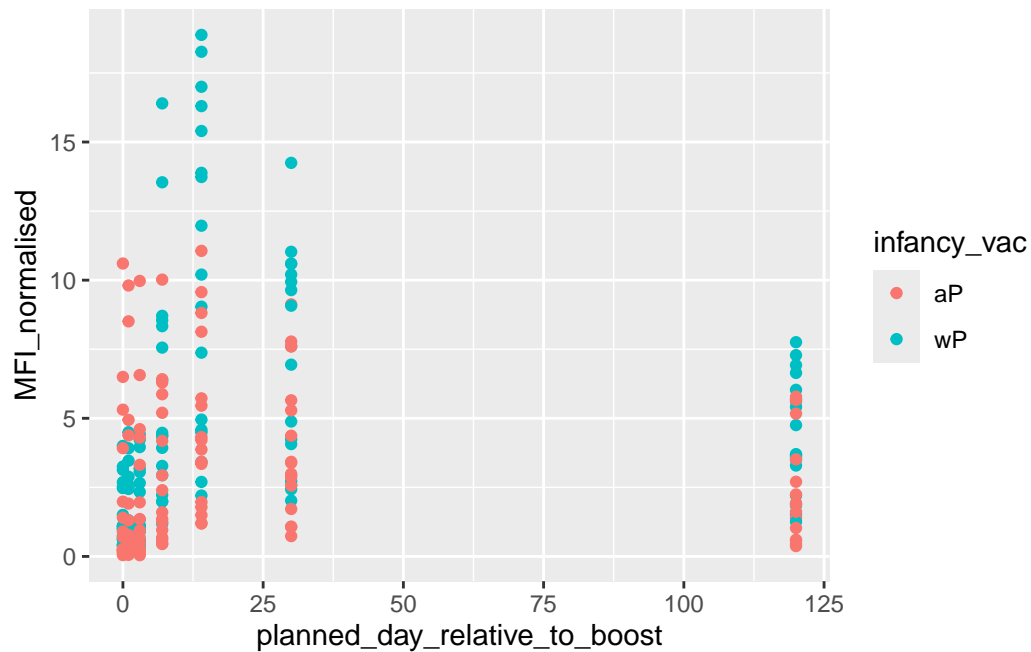
```
2021_dataset
8085
```

Focus on PT antigen for IgG levels

```
pt.21 <- filter(abdata.21, isotype == "IgG", antigen == "PT")
```

Plot of

```
ggplot(pt.21) +
  aes(x=planned_day_relative_to_boost,
      y=MFI_normalised,
      col=infancy_vac,) +
  geom_point()
```



```
ggplot(pt.21) +
  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", x = "Days After Boost", y = "MFI")
```



Q18. Do you see any clear difference in aP vs. wP responses?

wP responses show a greater peak at 14-days post-vaccination.