class17

Steven Gan

2022-11-23

Table of contents

Getting Started

```
vax <- read.csv("covid19vaccinesbyzipcode_test.csv")
```

Q1:

persons_fully_vaccinated

Q2:

zip_code_tabulation_area

Q3:

```
sort(vax$as_of_date)[1]
```

```
[1] "2021-01-05"
```

Q4:

```
sort(vax$as_of_date, decreasing = T)[1]
```

```
[1] "2022-11-22"
```

```
skimr::skim(vax)
```

Table 1: Data summary

Name	vax
Number of rows	174636
Number of columns	18
_____	
Column type frequency:	
character	5
numeric	13
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
as_of_date	0	1	10	10	0	99	0
local_health_jurisdiction	0	1	0	15	495	62	0
county	0	1	0	15	495	59	0
vem_source	0	1	15	26	0	3	0
redacted	0	1	2	69	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
zip_code_tabulation_area	0	1.00	93665.11	1817.39	90001	92257.75	93658.50	95380.50	97635.0	
vaccine_equity_metric_quartile	8613	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	
age12_plus_population	0	1.00	18895.01	18993.88	0	1346.95	13685.13	31756.12	88556.7	
age5_plus_population	0	1.00	20875.22	21105.98	0	1460.50	15364.00	34877.00	101902.0	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
tot_population	8514	0.95	23372.72	72628.51	12	2126.00	18714.00	68168.00	1111165.0	
persons_fully_vaccinated	14921	0.91	13466.34	44722.46	11	883.00	8024.00	22529.00	687186.0	
persons_partially_vaccinated	14921	0.91	1707.50	1998.80	11	167.00	1194.00	2547.00	39204.0	
percent_of_population_fully_vaccinated	18665	0.89	0.55	0.25	0	0.39	0.59	0.73	1.0	
percent_of_population_partially_vaccinated	18665	0.89	0.08	0.09	0	0.05	0.06	0.08	1.0	
percent_of_population_with_1_plus_dose	19762	0.89	0.61	0.25	0	0.46	0.65	0.79	1.0	
booster_recip_count	70421	0.60	5655.17	6867.49	11	280.00	2575.00	9421.00	58304.0	
bivalent_dose_recip_count	156958	0.10	1646.02	2161.84	11	109.00	719.00	2443.00	18109.0	
eligible_recipient_count	0	1.00	12309.11	34555.83	0	466.00	5810.00	21140.00	866696.0	

Q5:

13

Q6:

```
sum(is.na(vax$persons_fully_vaccinated))
```

```
[1] 14921
```

Q7:

```
per <- round(sum(is.na(vax$persons_fully_vaccinated)) / nrow(vax) * 100, 1)
paste0(per, "%")
```

```
[1] "8.5%"
```

Q8:

Possibly due to that data remained uncollected

Working with dates

```
library(lubridate)
```

```
Loading required package: timechange
```

```
Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

```
today()
```

```
[1] "2022-11-23"
```

```
vax$as_of_date <- ymd(vax$as_of_date)

vax$as_of_date[nrow(vax)] - vax$as_of_date[1]
```

```
Time difference of 686 days
```

Q9:

```
today() - sort(vax$as_of_date, decreasing = T)[1]
```

```
Time difference of 1 days
```

Q10:

```
length(unique(vax$as_of_date))
```

```
[1] 99
```

Working with ZIP codes

```
# install.packages("zipcodeR")
library(zipcodeR)
```

```
geocode_zip('92037')
```

```
# A tibble: 1 x 3
  zipcode   lat   lng
  <chr>   <dbl> <dbl>
1 92037    32.8 -117.
```

```
zip_distance('92037','92109')
```

```
  zipcode_a zipcode_b distance
1     92037     92109     2.33
```

```
reverse_zipcode(c('92037', "92109"))
```

```
# A tibble: 2 x 24
  zipcode zipcode_~1 major~2 post_~3 common_c~4 county state   lat   lng timez~5
  <chr>   <chr>      <chr>   <chr>       <blob> <chr>  <chr> <dbl> <dbl> <chr>
1 92037   Standard   La Jol~ La Jol~ <raw 20 B> San D~ CA     32.8 -117. Pacific
2 92109   Standard   San Di~ San Di~ <raw 21 B> San D~ CA     32.8 -117. Pacific
# ... with 14 more variables: radius_in_miles <dbl>, area_code_list <blob>,
#   population <int>, population_density <dbl>, land_area_in_sqmi <dbl>,
#   water_area_in_sqmi <dbl>, housing_units <int>,
#   occupied_housing_units <int>, median_home_value <int>,
#   median_household_income <int>, bounds_west <dbl>, bounds_east <dbl>,
#   bounds_north <dbl>, bounds_south <dbl>, and abbreviated variable names
#   1: zipcode_type, 2: major_city, 3: post_office_city, ...
```

```
zipdata <- reverse_zipcode(vax$zip_code_tabulation_area)
```

OPTION: Plotting on the percentage of fully vacination

```
# install.packages("maps")
library(maps)
library(ggplot2)
```

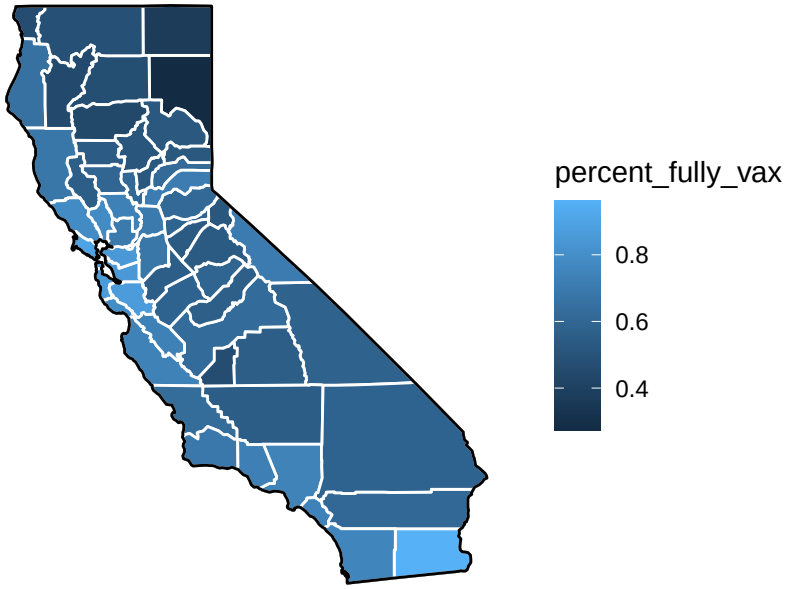
```
states <- map_data("state")
ca_df <- subset(states, region == "california")
counties <- map_data("county")
```

```r
ca_county <- subset(counties, region == "california")
```

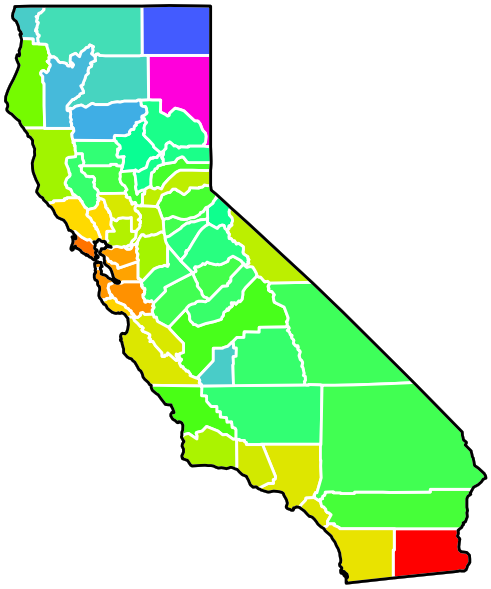
```r
library(stringr)
library(dplyr)
```

```r
vax_county <- vax %>%
  filter(as_of_date == "2022-11-22") %>%
  group_by(county) %>%
  summarise(per = sum(persons_fully_vaccinated, na.rm = T) /
              sum(tot_population, na.rm = T)) %>%
  na.omit() %>%
  `colnames<-`(c("subregion", "percent_fully_vax")) %>%
  mutate_if(is.character, str_to_lower)

cacopa <- inner_join(ca_county, vax_county, by = "subregion")
```

```r
ditch_the_axes <- theme(
  axis.text = element_blank(),
  axis.line = element_blank(),
  axis.ticks = element_blank(),
  panel.border = element_blank(),
  panel.grid = element_blank(),
  axis.title = element_blank()
  )

ca_base <- ggplot(data = ca_df, mapping = aes(x = long, y = lat, group = group)) +
  coord_fixed(1.3) +
  geom_polygon(color = "black", fill = "gray")

elbow_room1 <- ca_base +
      geom_polygon(data = cacopa, aes(fill = percent_fully_vax), color = "white") +
      geom_polygon(color = "black", fill = NA) +
      theme_bw() +
      ditch_the_axes

elbow_room1
```




```
eb2 <- elbow_room1 +
    scale_fill_gradientn(colours = rev(rainbow(7)),
                         breaks = c(2, 4, 10, 100, 1000, 10000),
                         trans = "log10")
eb2
```



Focus on the San Diego area

```
sd <- vax %>% filter(county == "San Diego")
nrow(sd)
```

```
[1] 10593
```

```
sd.10 <- vax %>%
  filter(county == "San Diego" & age5_plus_population > 10000)
```

Q11:

```
sd %>%
  pull(zip_code_tabulation_area) %>%
  unique() %>%
  length()
```

```
[1] 107
```

Q12:

```
sd %>%
  arrange(desc(age12_plus_population)) %>%
  filter(row_number() == 1) %>%
  pull(zip_code_tabulation_area)
```

```
[1] 92154
```

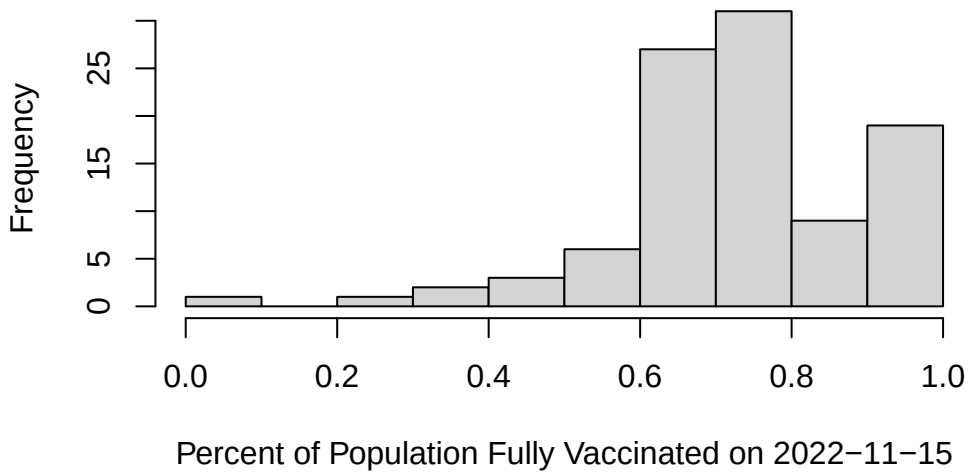
Q13:

```
sd_20221115 <- sd %>% filter(as_of_date == ymd("2022-11-15"))

sd_20221115 %>%
  summarise(per = sum(persons_fully_vaccinated, na.rm = T) /
              sum(tot_population, na.rm = T) * 100)
```

```
      per
1 75.2562
```

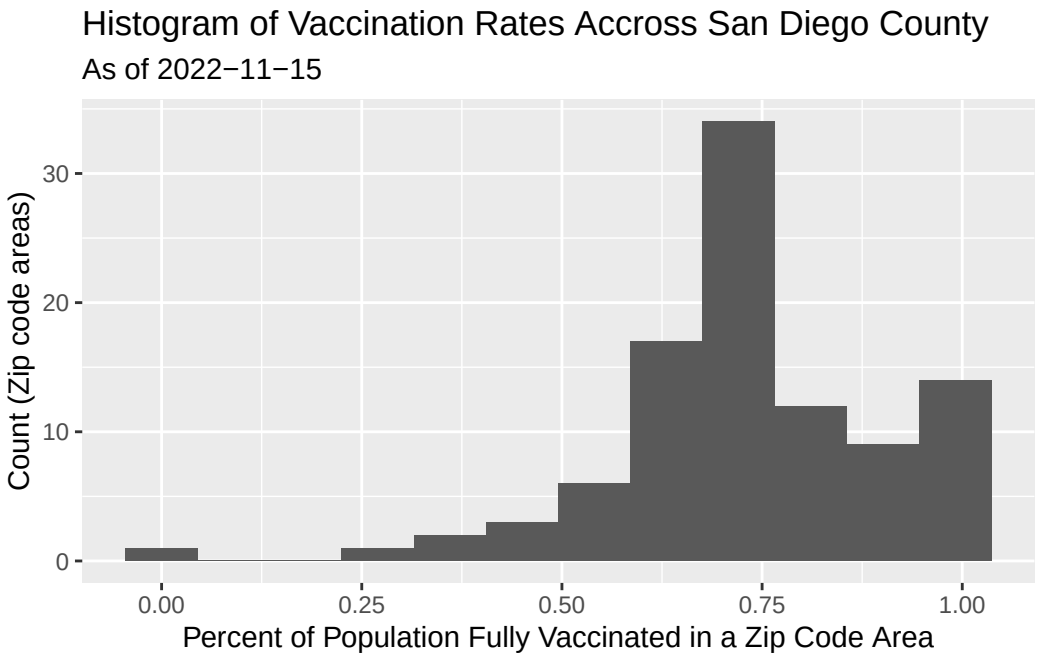
Q14:

```
hist(sd_20221115$percent_of_population_fully_vaccinated,
     xlab = "Percent of Population Fully Vaccinated on 2022-11-15",
     main = "Histogram of Vaccination Rates Accross San Diego County")
```




```
ggplot(sd_20221115) + aes(percent_of_population_fully_vaccinated) +
  geom_histogram(bins = 12) +
  labs(x = "Percent of Population Fully Vaccinated in a Zip Code Area",
       y = "Count (Zip code areas)",
       title = "Histogram of Vaccination Rates Accross San Diego County",
       subtitle = "As of 2022-11-15")
```

```
Warning: Removed 8 rows containing non-finite values (`stat_bin()`).
```



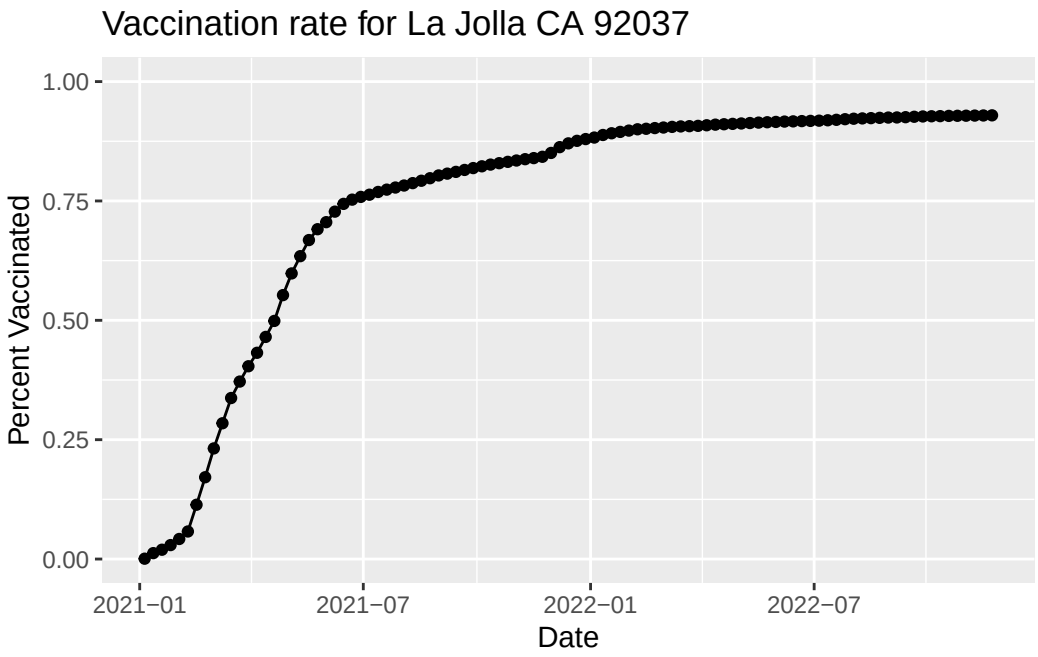
Focus on UCSD/La Jolla

```
ucsd <- sd %>%
  filter(zip_code_tabulation_area == "92037")
ucsd[1,]$age5_plus_population
```

```
[1] 36144
```

Q15:

```
a <- ggplot(ucsd) +
  aes(as_of_date, percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line() +
  ylim(c(0, 1)) +
  labs(x = "Date", y = "Percent Vaccinated",
       title = "Vaccination rate for La Jolla CA 92037")

a
```

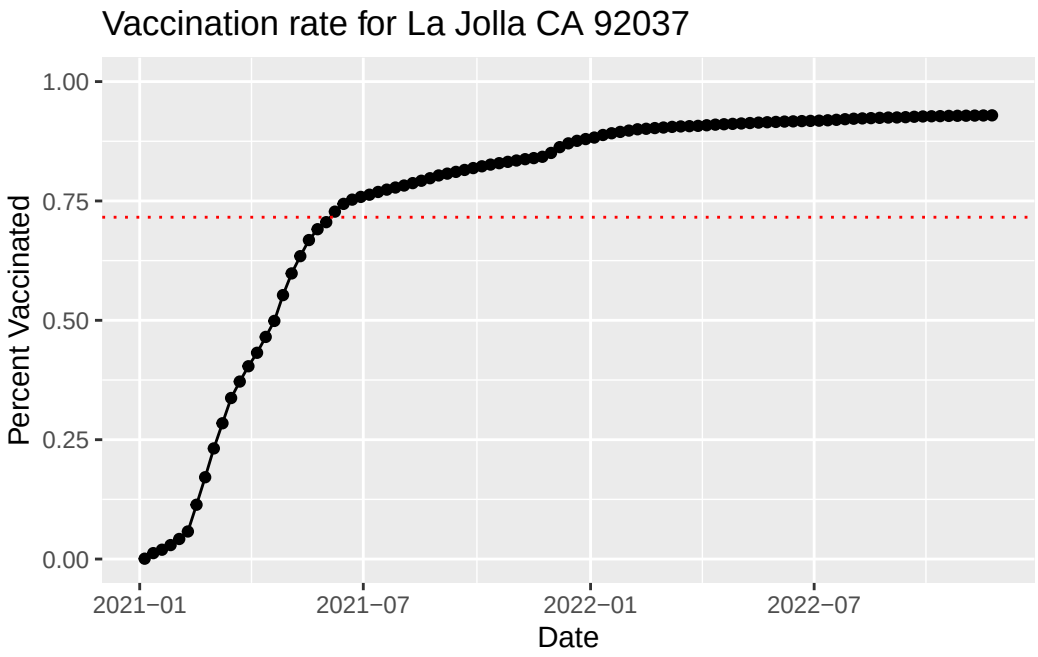


Comparing to similar sized areas

```
vax.36 <- vax %>%
  filter(age5_plus_population > 36144 & as_of_date == "2022-11-15")
```

Q16:

```
mean_per_full <- vax.36 %>%
  summarise(Mean_per_full = sum(persons_fully_vaccinated) / sum(tot_population))

a + geom_hline(yintercept = as.numeric(mean_per_full),
               linetype = "dotted", col = "red")
```



Q17:

```
vax.36 %>%
  summarise(Min = min(percent_of_population_fully_vaccinated),
            fQu = quantile(percent_of_population_fully_vaccinated, probs = c(0.25)),
            Median = median(percent_of_population_fully_vaccinated),
            Mean = mean(percent_of_population_fully_vaccinated),
            tQu = quantile(percent_of_population_fully_vaccinated, probs = c(0.75)),
            Max = max(percent_of_population_fully_vaccinated))
```

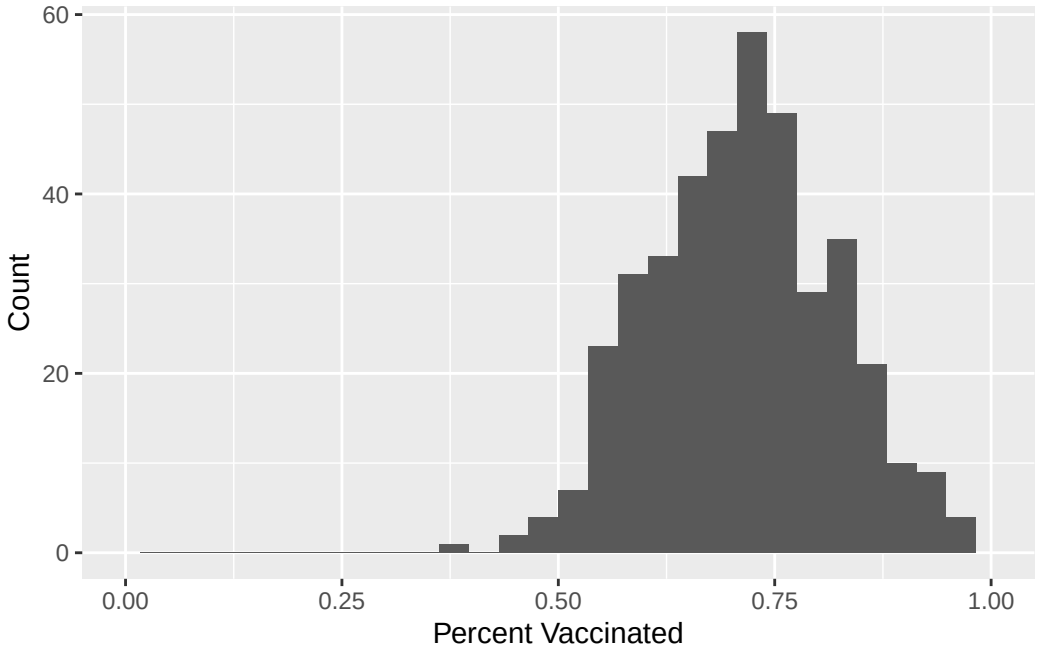
```
       Min       fQu   Median      Mean      tQu Max
1 0.378501 0.6396185 0.715524 0.7172851 0.787982   1
```

Q18:

```
ggplot(vax.36) + aes(percent_of_population_fully_vaccinated) +
  geom_histogram() +
  xlim(c(0, 1)) +
  labs(x = "Percent Vaccinated", y = "Count")
```

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

```
Warning: Removed 2 rows containing missing values (`geom_bar()`).
```



Q19:

```
vax %>%
  filter(as_of_date == "2022-11-15") %>%
  filter(zip_code_tabulation_area == "92040") %>%
  select(percent_of_population_fully_vaccinated)
```

```
  percent_of_population_fully_vaccinated
1                               0.546646
```

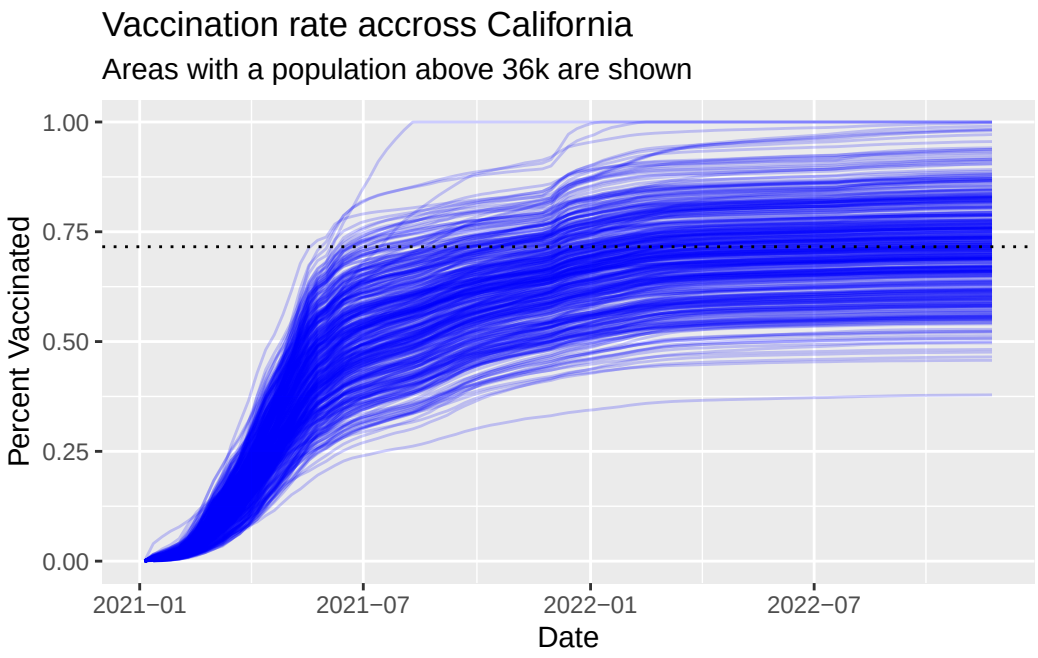
Below the average value

Q20:

```
vax.36.all <- vax %>%
  filter(age5_plus_population > 36144)

ggplot(vax.36.all) + aes(as_of_date, percent_of_population_fully_vaccinated,
                         group = zip_code_tabulation_area) +
```

```
    geom_line(alpha = 0.2, col = "blue") +
    ylim(c(0, 1)) +
    labs(x = "Date", y = "Percent Vaccinated",
         title = "Vaccination rate accross California",
         subtitle = "Areas with a population above 36k are shown") +
    geom_hline(yintercept = as.numeric(mean_per_full), linetype = "dotted")
```

```
Warning: Removed 184 rows containing missing values (`geom_line()`).
```



Q21:

Highly agree!