Lab14: COVID-19 vaccination rate mini project

Rachel Kraft

2022-03-03

read data set for vaccination rates

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

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction         county
## 1 2021-01-05                   92549                 Riverside      Riverside
## 2 2021-01-05                   92130                 San Diego      San Diego
## 3 2021-01-05                   92397           San Bernardino San Bernardino
## 4 2021-01-05                   94563             Contra Costa   Contra Costa
## 5 2021-01-05                   94519             Contra Costa   Contra Costa
## 6 2021-01-05                   91042              Los Angeles    Los Angeles
##   vaccine_equity_metric_quartile               vem_source
## 1                              3 Healthy Places Index Score
## 2                              4 Healthy Places Index Score
## 3                              3 Healthy Places Index Score
## 4                              4 Healthy Places Index Score
## 5                              3 Healthy Places Index Score
## 6                              2 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1                2348.4                 2461                       NA
## 2               46300.3                53102                       61
## 3                3695.6                 4225                       NA
## 4               17216.1                18896                       NA
## 5               16861.2                18678                       NA
## 6               23962.2                25741                       NA
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                           NA                                     NA
## 2                           27                                0.001149
## 3                           NA                                     NA
## 4                           NA                                     NA
## 5                           NA                                     NA
## 6                           NA                                     NA
##   percent_of_population_partially_vaccinated
## 1                                         NA
## 2                                   0.000508
## 3                                         NA
## 4                                         NA
## 5                                         NA
## 6                                         NA
##   percent_of_population_with_1_plus_dose booster_recip_count
```

```
## 1                                     NA                   NA
## 2                                0.001657                  NA
## 3                                     NA                   NA
## 4                                     NA                   NA
## 5                                     NA                   NA
## 6                                     NA                   NA
##                                                                 redacted
## 1 Information redacted in accordance with CA state privacy requirements
## 2 Information redacted in accordance with CA state privacy requirements
## 3 Information redacted in accordance with CA state privacy requirements
## 4 Information redacted in accordance with CA state privacy requirements
## 5 Information redacted in accordance with CA state privacy requirements
## 6 Information redacted in accordance with CA state privacy requirements
```

Q1. What column details the total number of people fully vaccinated?

persons_fully_vaccinated

Q2. What column details the Zip code tabulation area?

zip_code_tabulation_area

Q3. What is the earliest date in this dataset?

2021-01-05

Q4. What is the latest date in this dataset?

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

```
## [1] "2022-03-01"
```

2022-03-01

call the skim() function from the skimr package to get a quick overview of this dataset

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

```
skim(vax)
```

Table 1: Data summary

Name	vax
Number of rows	107604
Number of columns	15
Column type frequency:	
character	5

Table 1: Data summary

numeric	10
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	61	0
local_health_jurisdiction	0	1	0	15	305	62	0
county	0	1	0	15	305	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	5307	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	
age12_plus_population	0	1.00	18895.04	18993.91	0	1346.95	13685.10	31756.12	88556.7	
age5_plus_population	0	1.00	20875.24	21106.02	0	1460.50	15364.00	34877.00	101902.0	
persons_fully_vaccinated	18338	0.83	12155.61	13063.88	11	1066.25	7374.50	20005.00	77744.0	
persons_partially_vaccinated	18338	0.83	831.74	1348.68	11	76.00	372.00	1076.00	34219.0	
percent_of_population_fully_vaccinated	18338	0.83	0.51	0.26	0	0.33	0.54	0.70	1.0	
percent_of_population_partially_vaccinated	18338	0.83	0.05	0.09	0	0.01	0.03	0.05	1.0	
percent_of_population_with_1_plus_dose	18338	0.83	0.54	0.28	0	0.36	0.58	0.75	1.0	
booster_recip_count	64317	0.40	4100.55	5900.21	11	176.00	1136.00	6154.50	50602.0	

Q5. How many numeric columns are in this dataset?

9

Q6. Note that there are "missing values" in the dataset. How many NA values there in the persons_fully_vaccinated column?

18338

Q7. What percent of persons_fully_vaccinated values are missing (to 2 significant figures)?

```
round((18338/107604*100),2)
```

```
## [1] 17.04
```

Working with Dates

we can use the lubridate package to more easily work with dates

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

```
age <- today() - ymd("2000-07-18")
age
```

```
## Time difference of 7898 days
```

```
time_length(age, "year")
```

```
## [1] 21.62355
```

convert our date data into a lubridate format

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

how many days since the first vaccination reported in the data set?

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

```
## Time difference of 422 days
```

Q9. How many days have passed since the last update of the dataset?

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

```
## Time difference of 2 days
```

Q10. How many unique dates are in the dataset?

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

```
## [1] 61
```

Working with zip codes

use the zipcodeR package

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

locate la jolla

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

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

calculate the distance between two zipcodes

```
zip_distance('92037','91362')
```

```
##   zipcode_a zipcode_b distance
## 1     92037     91362   134.38
```

use reverse_zipcode() to pull census data

Focus on San Diego Area

restrict to vax$county == “San Diego” entries

using r

```
# Subset to San Diego county only areas
sd <- vax[2,]
```

```
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
```

```
sd <- filter(vax, county == "San Diego")

nrow(sd)
```

```
## [1] 6527
```

Q11. How many distinct zip codes are listed for San Diego County?

```
length(unique(sd$zip_code_tabulation_area))
```

```
## [1] 107
```

Q12. What San Diego County Zip code area has the largest 12 + Population in this dataset?

```
pop <- which.max(sd$age12_plus_population)
sd$zip_code_tabulation_area[pop]
```

```
## [1] 92154
```

Using dplyr select all San Diego “county” entries on “as_of_date” “2022-02-22”

```
feb <- filter(sd, as_of_date=="2022-02-22")
```

Q13. What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2022-02-22"?

```
sd$as_of_date[nrow(sd)]
```

```
## [1] "2022-03-01"
```

let's work with the most recent data

```
latest <- filter(sd, as_of_date=="2022-03-01")
mean(latest$percent_of_population_fully_vaccinated, na.rm=TRUE)
```

```
## [1] 0.7052904
```

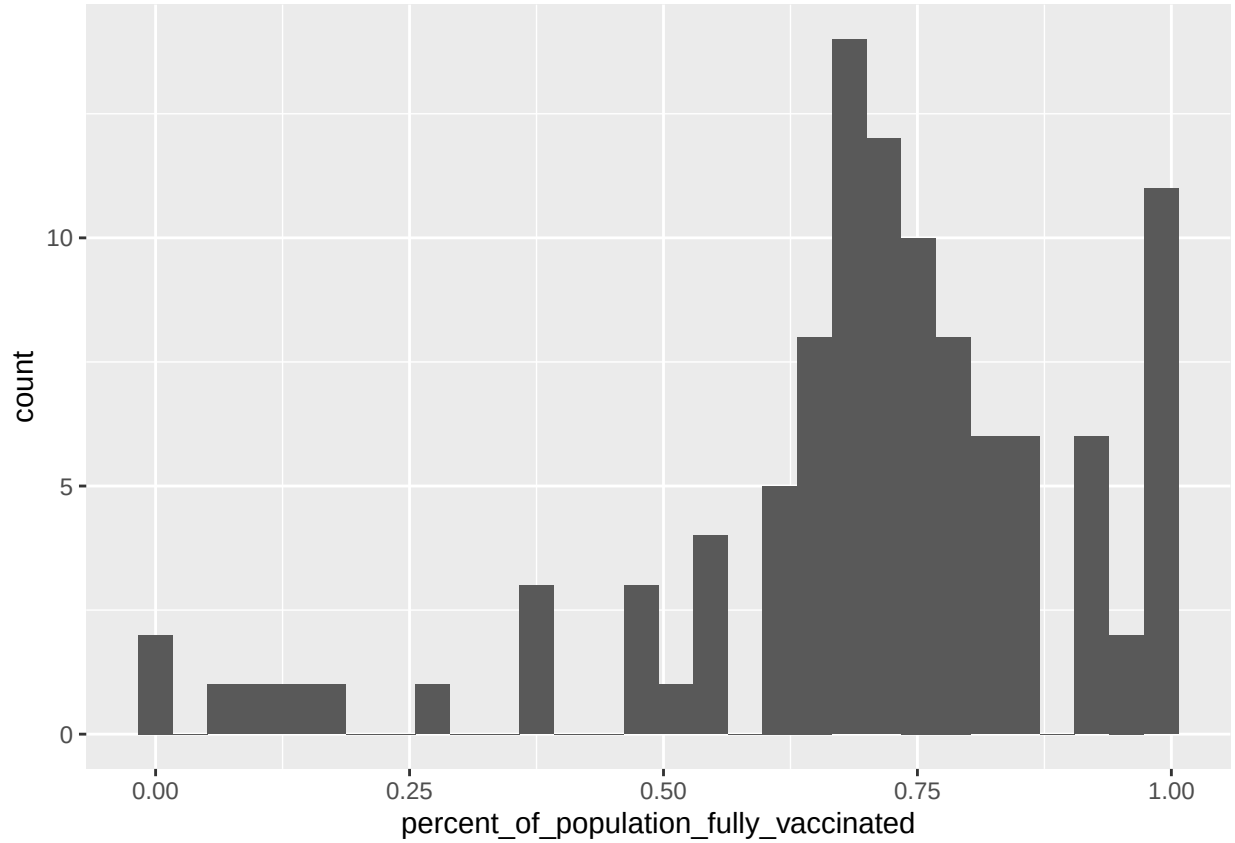
Q14. Using either ggplot or base R graphics make a summary figure that shows the distribution of Percent of Population Fully Vaccinated values as of "2022-03-01"?

```
library(ggplot2)
```

```
ggplot(latest)+aes(percent_of_population_fully_vaccinated)+geom_histogram()
```

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

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



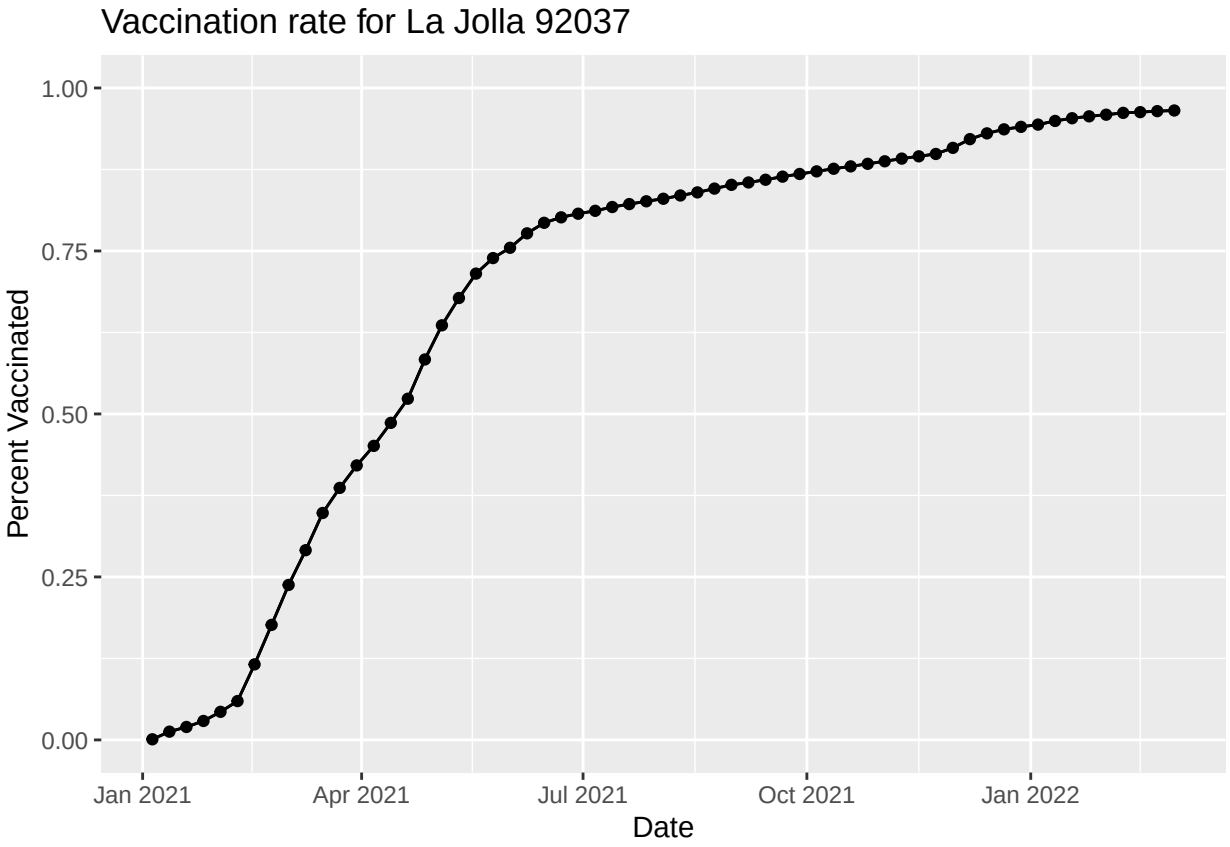
Focus on UCSD/La Jolla

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

```
## [1] 36144
```

Q15. Using ggplot make a graph of the vaccination rate time course for the 92037 ZIP code area

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



Comparing to similar sized areas

```
# Subset to all CA areas with a population as large as 92037
vax.36 <- filter(vax, age5_plus_population > 36144 &
                 as_of_date == "2022-03-01")
```

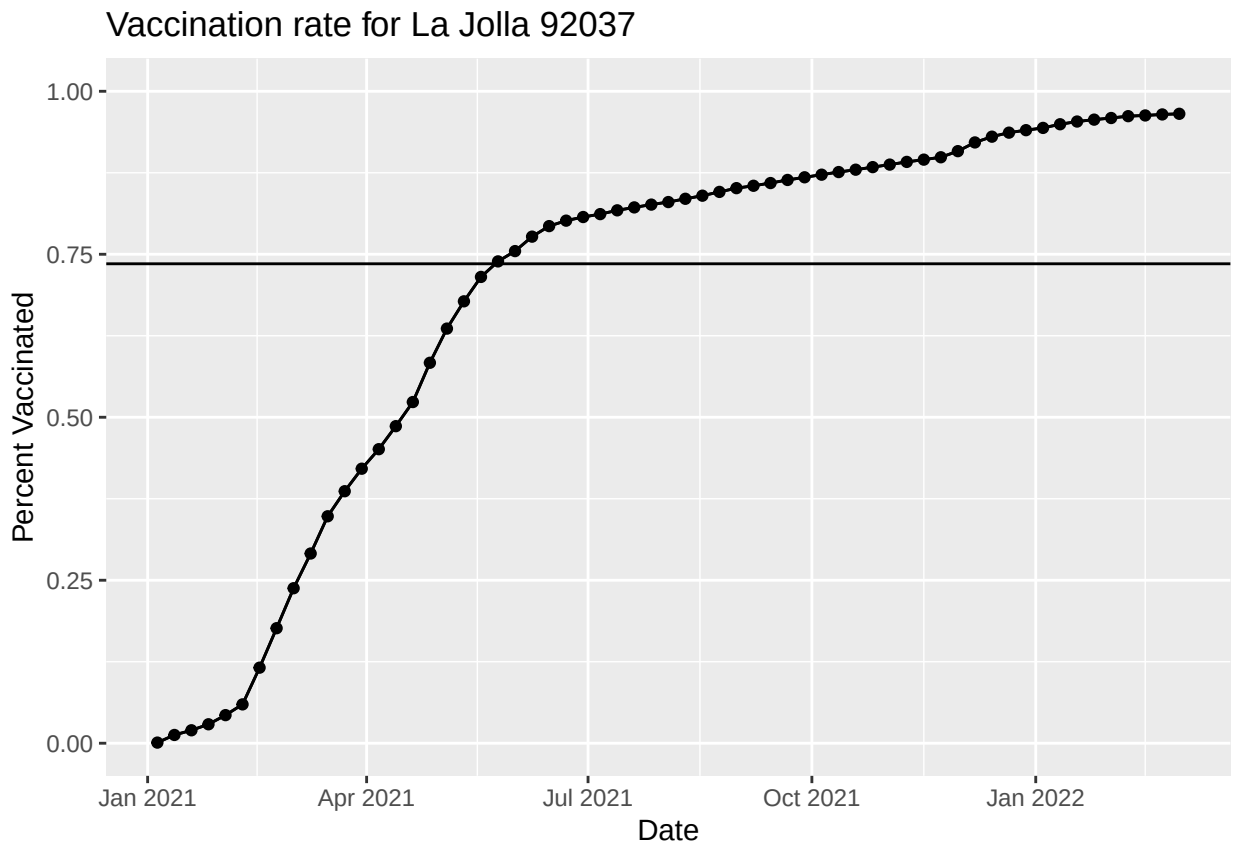
```
#head(vax.36)
```

Q16. Calculate the mean "Percent of Population Fully Vaccinated" for ZIP code areas with a population as large as 92037 (La Jolla) as_of_date "2022-03-01". Add this as a straight horizontal line to your plot from above with the geom_hline() function?

```
ave.36 <- mean(vax.36$percent_of_population_fully_vaccinated, na.rm=TRUE)
ave.36
```

```
## [1] 0.7353974
```

```
baseplot+geom_hline(yintercept=ave.36)
```



Q17. What is the 6 number summary (Min, 1st Qu., Median, Mean, 3rd Qu., and Max) of the "Percent of Population Fully Vaccinated" values for ZIP code areas with a population as large as 92037 (La Jolla) as_of_date "2022-03-01"?

```
summary(vax.36$percent_of_population_fully_vaccinated)
```

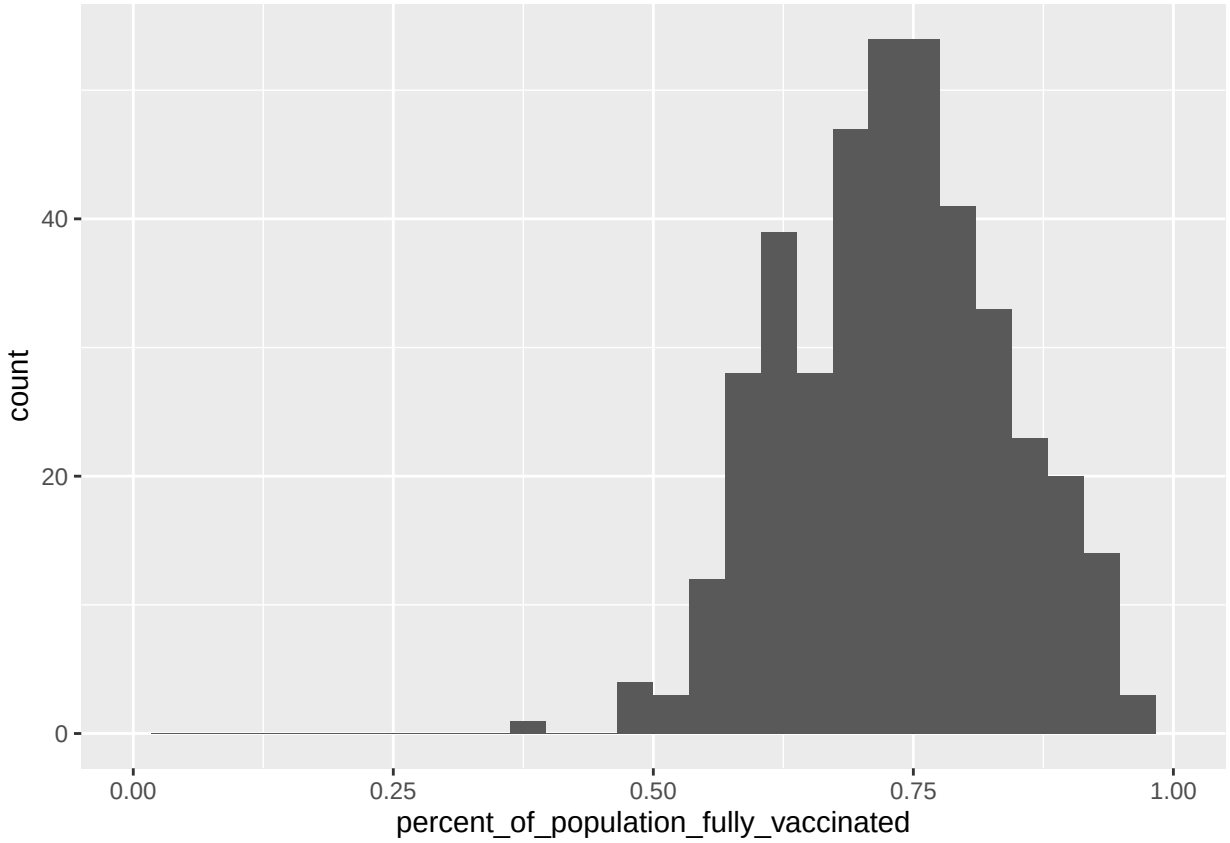
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3890  0.6554  0.7350  0.7354  0.8044  1.0000
```

Q18. Using ggplot generate a histogram of this data.

```
ggplot(vax.36) +
  aes(percent_of_population_fully_vaccinated) +
  geom_histogram() +
  xlim(c(0,1))
```

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

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



Q19. Is the 92109 and 92040 ZIP code areas above or below the average value you calculated for all these above?

```
vax %>% filter(as_of_date == "2022-03-01") %>%
  filter(zip_code_tabulation_area=="92040") %>%
  select(percent_of_population_fully_vaccinated)
```

```
##   percent_of_population_fully_vaccinated
## 1                               0.551981
```

```
vax %>% filter(as_of_date == "2022-03-01") %>%
  filter(zip_code_tabulation_area=="92109") %>%
  select(percent_of_population_fully_vaccinated)
```

```
##   percent_of_population_fully_vaccinated
## 1                               0.723778
```

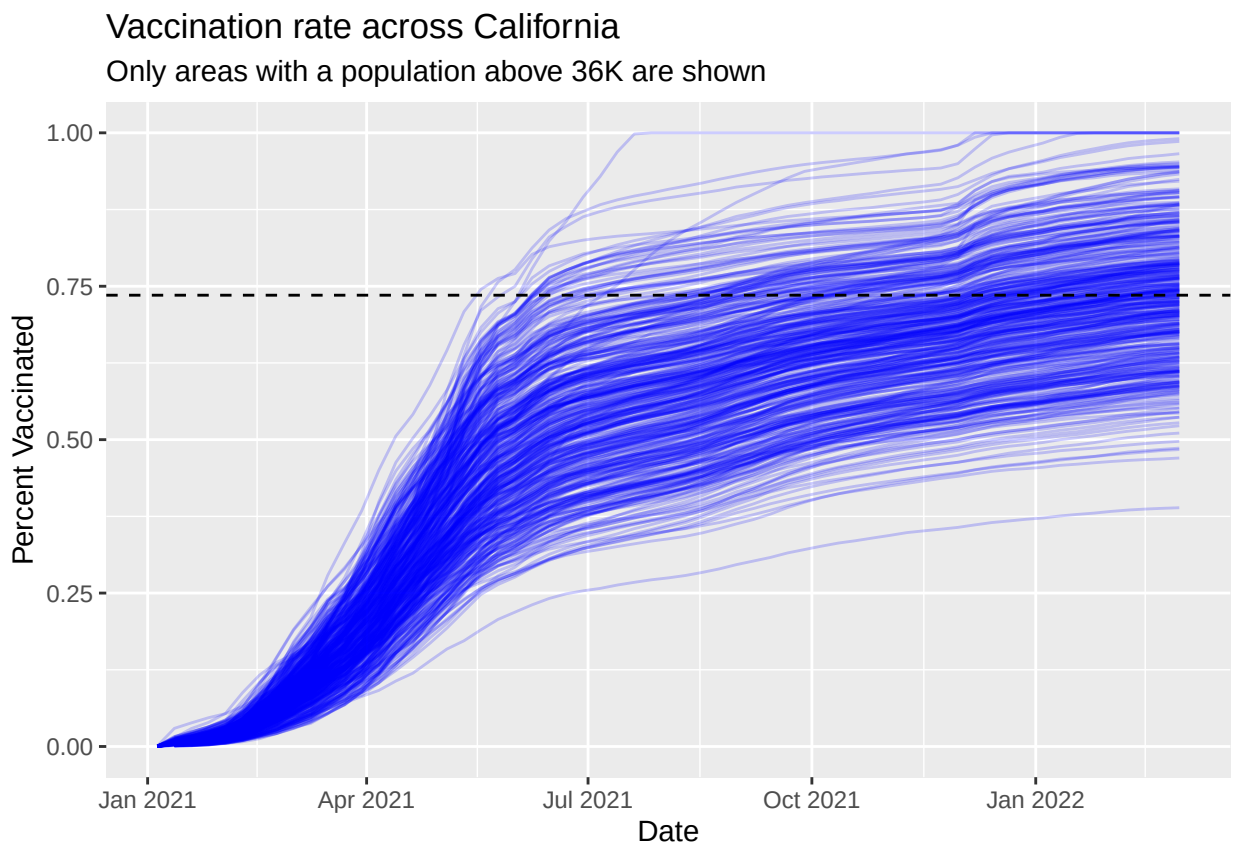
both areas have a lower fully vaccinated average

Q20. Finally make a time course plot of vaccination progress for all areas in the full dataset with a age5_plus_population > 36144

```
vax.36.all <- filter(vax, 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, color="blue") +
  ylim(c(0,1)) +
  labs(x="Date", y="Percent Vaccinated",
       title="Vaccination rate across California",
       subtitle="Only areas with a population above 36K are shown") +
  geom_hline(yintercept = ave.36, linetype=2)
```

```
## Warning: Removed 311 row(s) containing missing values (geom_path).
```



> 21. How do you feel about traveling for Spring Break and meeting for in-person class afterwards?

I think it is okay as long as everyone is testing after they come back, but I doubt that most people would do that so it makes me question if we should go back in person immediately. However, I don't think classes should be online anymore, that if everyone is wearing masks inside it should be fine.