Class 17: Vaccination Mini-Project

Garrett Cole

Getting Started

```
#import vaccination data
vax <- read.csv("covid19vaccinesbyzipcode_test.csv")
head(vax)
```

```
  as_of_date zip_code_tabulation_area local_health_jurisdiction          county
1 2021-01-05                   92240                 Riverside       Riverside
2 2021-01-05                   91302               Los Angeles     Los Angeles
3 2021-01-05                   93420           San Luis Obispo San Luis Obispo
4 2021-01-05                   91901                 San Diego       San Diego
5 2021-01-05                   94110             San Francisco   San Francisco
6 2021-01-05                   91902                 San Diego       San Diego
  vaccine_equity_metric_quartile                vem_source
1                              1 Healthy Places Index Score
2                              4 Healthy Places Index Score
3                              3 Healthy Places Index Score
4                              3 Healthy Places Index Score
5                              4 Healthy Places Index Score
6                              4 Healthy Places Index Score
  age12_plus_population age5_plus_population tot_population
1               29270.5                33093          35278
2               23163.9                25899          26712
3               26694.9                29253          30740
4               15549.8                16905          18162
5               64350.7                68320          72380
6               16620.7                18026          18896
  persons_fully_vaccinated persons_partially_vaccinated
1                       NA                           NA
2                       15                          614
3                       NA                           NA
```

```
4                       NA                            NA
5                       17                          1268
6                       15                           397
  percent_of_population_fully_vaccinated
1                                     NA
2                                0.000562
3                                     NA
4                                     NA
5                                0.000235
6                                0.000794
  percent_of_population_partially_vaccinated
1                                         NA
2                                    0.022986
3                                         NA
4                                         NA
5                                    0.017519
6                                    0.021010
  percent_of_population_with_1_plus_dose booster_recip_count
1                                     NA                  NA
2                                0.023548                 NA
3                                     NA                  NA
4                                     NA                  NA
5                                0.017754                 NA
6                                0.021804                 NA
  bivalent_dose_recip_count eligible_recipient_count
1                        NA                       2
2                        NA                      15
3                        NA                       4
4                        NA                       8
5                        NA                      17
6                        NA                      15
                                                                    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?

10

```
#Commented out for pdf space
#vax["persons_fully_vaccinated"]
#vax[10]
```

Question 2: What column details the Zip code tabulation area?

2

```
#Commented out for pdf space
#vax["zip_code_tabulation_area"]
#vax[2]
```

```
# dimensions of dataset
dim(vax)
```

```
[1] 174636     18
```

Question 3: What is the earliest date in this dataset?

2021-01-05

```
vax$as_of_date[1]
```

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

Question 4: What is the latest date in this dataset?

2022-11-22

```
vax$as_of_date[174636]
```

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

```
# Overview of vax
library(skimr)
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	18895.04	18993.88	0	1346.95	13685.50	31756.25	88556.7	
age5_plus_population	0	1.00	20875.22	21105.98	0	1460.50	15364.00	34877.00	101902.0	
tot_population	8514	0.95	23372.72	22628.51	12	2126.00	18714.00	38168.00	111165.0	
persons_fully_vaccinated	14921	0.91	13466.34	14722.46	11	883.00	8024.00	22529.00	67186.0	
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	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
percent_of_population_with_1_plus_dose	19562	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	14555.83	0	466.00	5810.00	21140.00	86696.0	

Question 5: How many numeric columns are in this dataset?

There are 9 numeric columns # Question 6: Note that there are "missing values" in the dataset. How many NA values there in the persons_fully_vaccinated column? 14921

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

```
[1] 14921
```

Question 7: What percent of persons_fully_vaccinated values are missing (to 2 significant figures)

8.54%

```
#convert NA values to 0
vax[is.na(vax)] = 0
```

```
#double check number of missing values is 14921
colSums(vax==0)
```

```
                            as_of_date
                                     0
                 zip_code_tabulation_area
                                     0
               local_health_jurisdiction
                                     0
                                county
                                     0
          vaccine_equity_metric_quartile
                                  8613
                            vem_source
```

```
                                     0
                       age12_plus_population
                                  2574
                        age5_plus_population
                                  2574
                             tot_population
                                  8514
                    persons_fully_vaccinated
                                 14921
                persons_partially_vaccinated
                                 14921
    percent_of_population_fully_vaccinated
                                 18665
percent_of_population_partially_vaccinated
                                 18665
    percent_of_population_with_1_plus_dose
                                 19562
                         booster_recip_count
                                 70421
                  bivalent_dose_recip_count
                                156958
                   eligible_recipient_count
                                  1499
                               redacted
                                     0
```

```
# divide num. of missing values by total num. of values to find percent missing
(14921/174636) * 100
```

```
[1] 8.544057
```

Question 8 (Optional): Why might this data be missing?

This data might be missing because there wasn't a confirmed number of persons fully vaccinated for that entry at the time ## 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] "2022-11-28"
```

```
# Specify that we are using the year-month-day format
vax$as_of_date <- ymd(vax$as_of_date)
```

```
# time since first date to today
today() - vax$as_of_date[1]
```

```
Time difference of 692 days
```

```
# time since first date to last date
vax$as_of_date[nrow(vax)] - vax$as_of_date[1]
```

```
Time difference of 686 days
```

Question 9: How many days have passed since the last update of the dataset?

6 days since last update

```
# time since last update to today's date
today() - vax$as_of_date[nrow(vax)]
```

```
Time difference of 6 days
```

Question 10: How many unique dates are in the dataset (i.e. how many different dates are detailed)?

99 unique dates

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

```
[1] 99
```

Working with Zip Codes

```
library(zipcodeR)
```

```
#centroid of La Jolla zip code 92037
geocode_zip('92037')
```

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

```
#distance between two centroids
zip_distance('92037','92109')
```

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

```
#census data about zip code areas
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, ...
```

Focus on San Diego Area

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

```
#using dplyr
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] 10593
```

```
# all San Diego county areas with a population of over 10,000
sd.10 <- filter(vax, county == "San Diego" &
                age5_plus_population > 10000)
```

Question 11: How many distinct zip codes are listed for San Diego County?

107 distinct zip codes

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

```
[1] 107
```

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

92154

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

```
[1] 92154
```

```
#select all San Diego "county" entries on "as_of_date" "2022-11-15"
tempSD <- filter(vax, county == "San Diego" &
                 as_of_date == "2022-11-15")
```

Question 13: What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2022-11-15"?

0.6818

```
mean(tempSD$percent_of_population_fully_vaccinated)
```

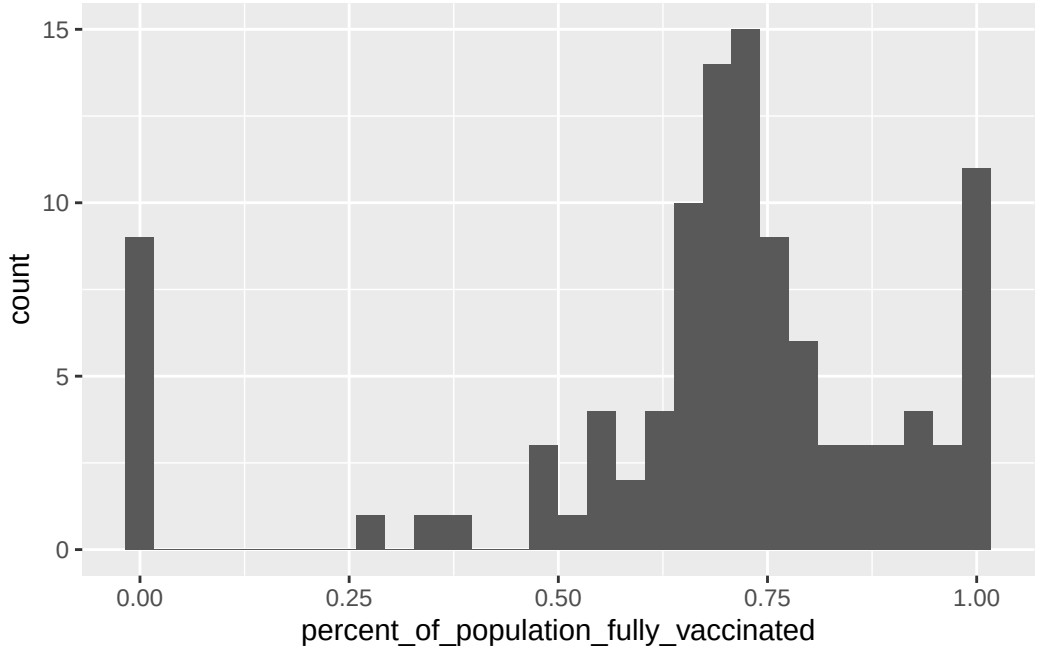
```
[1] 0.6818138
```

Question 14: Using either ggplot or base R graphics make a summary figure that shows the distribution of Percent of Population Fully Vaccinated values as of "2022-11-15"?

```
library(ggplot2)
```

```
ggplot(tempSD, aes(percent_of_population_fully_vaccinated)) + geom_histogram()
```

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



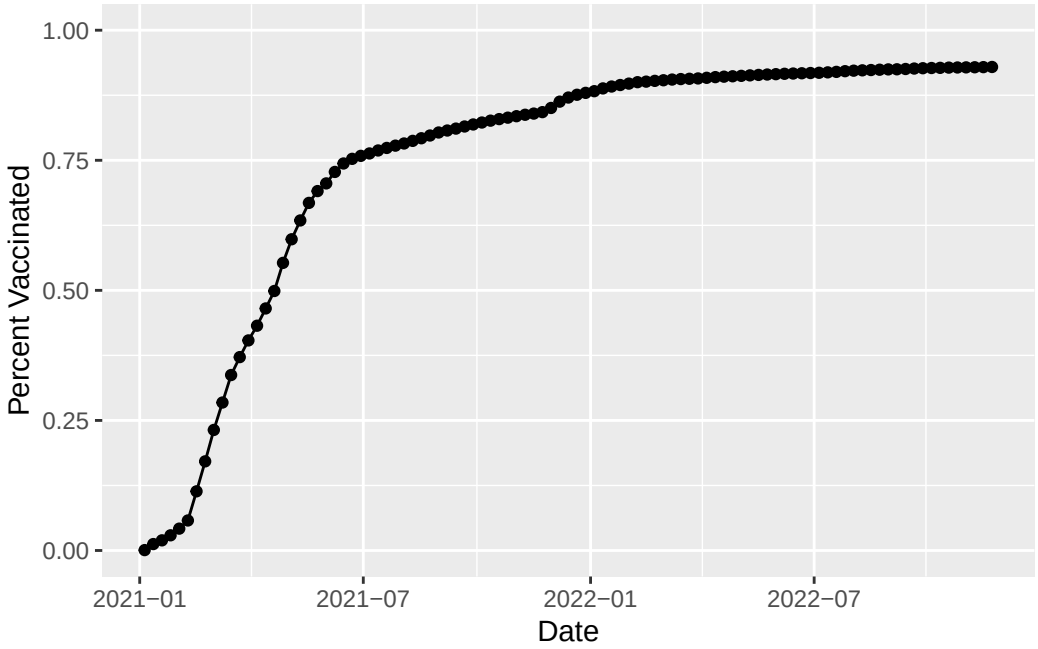
Focus on UCSD/La Jolla

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

```
[1] 36144
```

Question 15: Using ggplot make a graph of the vaccination rate time course for the 92037 ZIP code area

```
ggplot(ucsd) +
  aes(as_of_date,
      percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) +
  ylim(c(0,1)) +
  labs(x = "Date", y="Percent Vaccinated")
```



Comparing to similar sized areas

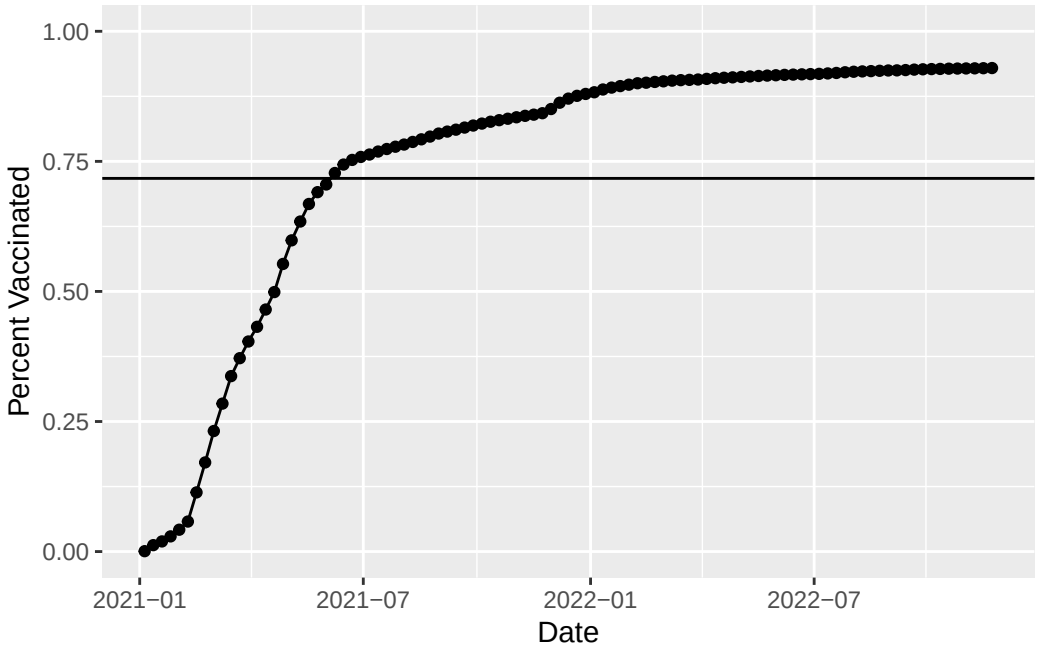
```
# Subset to all CA areas with a population as large as 92037
vax.36 <- filter(vax, age5_plus_population > 36144 &
                 as_of_date == "2022-11-15")

#head(vax.36)
```

Question 16: Calculate the mean "Percent of Population Fully Vaccinated" for ZIP code areas with a population as large as 92037 (La Jolla) as_of_date "2022-11-15". Add this as a straight horizontal line to your plot from above with the geom_hline() function?

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

```
[1] 0.7172851
```

```
#add geom_hline()
ggplot(ucsd) +
  aes(as_of_date,
      percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) +
  geom_hline(yintercept = 0.7172851) +
  ylim(c(0,1)) +
  labs(x = "Date", y="Percent Vaccinated")
```



Question 17: 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 "2022-11-15"?

Min: 0.3785010 1st Qu.: 0.6396185 Median: 0.7155240 Mean: 0.7172851 3rd Qu.: 0.7879820 Max: 1.000000

```
#min, 1st quarter, median, 3rd quarter, max
fivenum(vax.36$percent_of_population_fully_vaccinated)
```

```
[1] 0.3785010 0.6396185 0.7155240 0.7879820 1.0000000
```

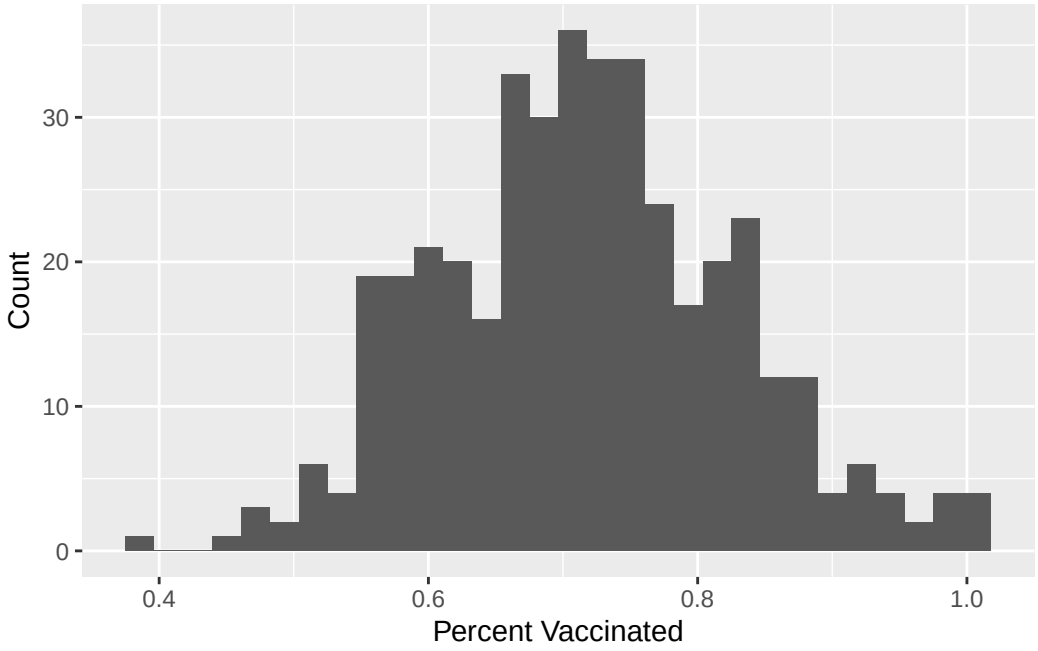
```
#mean
mean(vax.36$percent_of_population_fully_vaccinated)
```

```
[1] 0.7172851
```

Question 18: Using ggplot generate a histogram of this data.

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

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



Question 19: Is the 92109 and 92040 ZIP code areas above or below the average value you calculated for all these above?

Below, 92040 is 0.5467 and 92109 is 0.6933

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

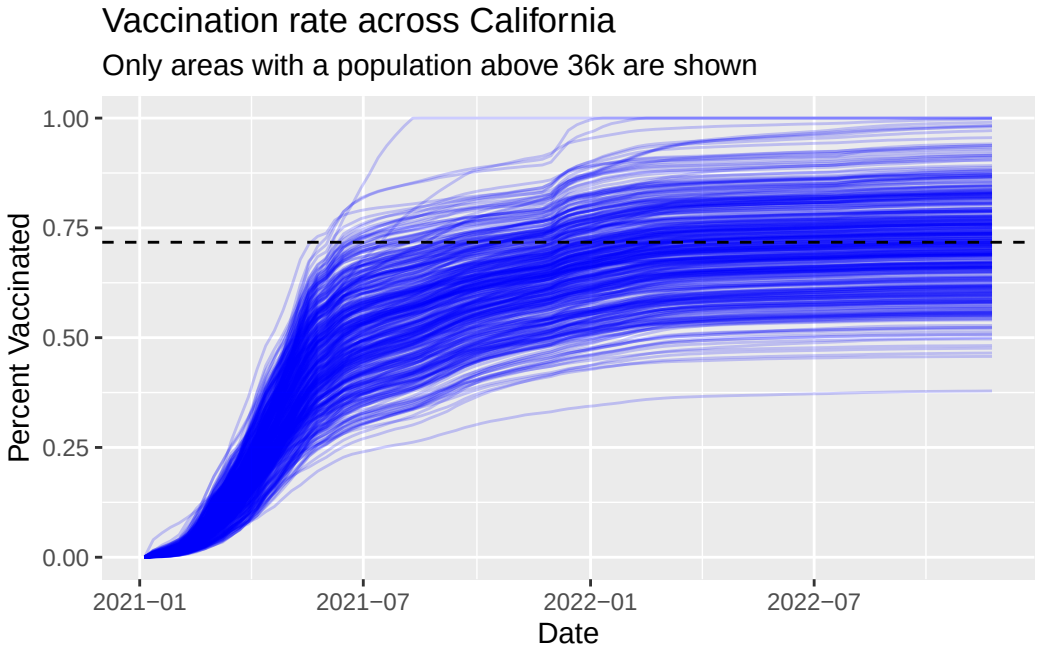
```
  percent_of_population_fully_vaccinated
1                               0.546646
```

```
#92109
vax %>% filter(as_of_date == "2022-11-15") %>%
  filter(zip_code_tabulation_area=="92109") %>%
  select(percent_of_population_fully_vaccinated)
```

```
  percent_of_population_fully_vaccinated
1                               0.693299
```

Question 20: 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 = 0.7172851, linetype = "dashed")
```



Question 21: How do you feel about traveling for Thanksgiving Break and meeting for in-person class afterwards?

I am excited to travel for Thanksgiving Break and looking forward to meeting for in-person class afterwards.