Vaccination Rate Miniproject

Liz Chamiec-Case

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?

"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?

“2022-11-22”

Q4. What is the latest date in this dataset?

“2021-01-05”

```
# skim dataset

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	93665.18	1817.39	90001	92257.75	93658.50	95380.00	97635.0	
vaccine_equity_metric_quartile	8613	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age12_plus_population	0	1.00	18895.08	93993.88	0	1346.95	3685.30	10756.8	2556.7	
age5_plus_population	0	1.00	20875.24	105105.98	0	1460.50	5364.50	4877.00	1902.0	
tot_population	8514	0.95	23372.27	7628.52	12	2126.00	8714.38	8168.00	1165.0	
persons_fully_vaccinated	14921	0.91	13466.34	722.46	16	883.00	8024.00	2529.87	7186.0	
persons_partially_vaccinated	14921	0.91	1707.50	998.80	1	167.00	1194.00	2547.00	9204.0	
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	19562	0.87	0.61	0.25	0	0.46	0.65	0.79	1.0	
booster_recip_count	70421	0.60	5655.16	7867.49	1	280.00	2575.00	9421.00	58304.0	
bivalent_dose_recip_count	156958	0.10	1646.02	2161.84	1	109.00	719.00	2443.00	8109.0	
eligible_recipient_count	0	1.00	12309.14	9555.80	3	466.00	5810.00	21140.00	6696.0	

Q5. How many numeric columns are in this dataset?

13

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

15440

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

0.09

Q8. [Optional]: Why might this data be missing?

Could be unaccounted for due to sources of vaccine not counted by this data (e.g. military, etc.)

Working with Dates

```
library(lubridate)
```

```
Warning: package 'lubridate' was built under R version 4.2.2
```

```
Loading required package: timechange
```

```
Warning: package 'timechange' was built under R version 4.2.2
```

```
Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

```
today()
```

```
[1] "2022-11-27"
```

```
# Specify that we are using the year-month-day format
vax$as_of_date <- ymd(vax$as_of_date)
today() - vax$as_of_date[1]
```

```
Time difference of 691 days
```

```
# date span of dataset

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

```
Time difference of 686 days
```

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

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

```
Time difference of 5 days
```

Q10. How many unique dates are in the dataset (i.e. how many different dates are detailed)?

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

```
[1] 99
```

Working with Zip Codes

```
library(zipcodeR)
```

```
Warning: package 'zipcodeR' was built under R version 4.2.2
```

```
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, ...
```

```
# Pull data for all ZIP codes in the dataset
# zipdata <- reverse_zipcode( vax$zip_code_tabulation_area ) # takes too long
```

San Diego Area

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

```
[1] 10593
```

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?

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

```
    as_of_date zip_code_tabulation_area local_health_jurisdiction    county
672 2021-01-05                   92154                San Diego San Diego
    vaccine_equity_metric_quartile                vem_source
672                              2 Healthy Places Index Score
    age12_plus_population age5_plus_population tot_population
672               76365.2                82971          88979
    persons_fully_vaccinated persons_partially_vaccinated
672                       17                         1379
    percent_of_population_fully_vaccinated
672                               0.000191
    percent_of_population_partially_vaccinated
672                                  0.015498
    percent_of_population_with_1_plus_dose booster_recip_count
672                               0.015689                  NA
    bivalent_dose_recip_count eligible_recipient_count
672                        NA                       17
                                                                redacted
672 Information redacted in accordance with CA state privacy requirements
```

92154

```r
# San Diego county on 2022-11-15

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
```

```r
sd.nov15 <- filter(vax, county == "San Diego", as_of_date == '2022-11-15')
head(sd.nov15)
```

```
  as_of_date zip_code_tabulation_area local_health_jurisdiction    county
1 2022-11-15                   92130                San Diego San Diego
2 2022-11-15                   91945                San Diego San Diego
3 2022-11-15                   92086                San Diego San Diego
4 2022-11-15                   92069                San Diego San Diego
5 2022-11-15                   92126                San Diego San Diego
6 2022-11-15                   92064                San Diego San Diego
  vaccine_equity_metric_quartile               vem_source
1                              4 Healthy Places Index Score
2                              2 Healthy Places Index Score
3                              1 Healthy Places Index Score
4                              2 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               46300.3                53102          56134
2               22820.5                25486          27236
3                1460.5                 1492           1543
4               41447.3                46850          50376
5               71820.2                77775          82658
6               42177.1                46855          49805
```

```
  persons_fully_vaccinated persons_partially_vaccinated
1                    52380                         5751
2                    19377                         1939
3                      761                           76
4                    34873                         2813
5                    60484                         5255
6                    36947                         2734
  percent_of_population_fully_vaccinated
1                                0.933124
2                                0.711448
3                                0.493195
4                                0.692254
5                                0.731738
6                                0.741833
  percent_of_population_partially_vaccinated
1                                   0.102451
2                                   0.071193
3                                   0.049255
4                                   0.055840
5                                   0.063575
6                                   0.054894
  percent_of_population_with_1_plus_dose booster_recip_count
1                               1.000000               34821
2                               0.782641               10425
3                               0.542450                 445
4                               0.748094               19456
5                               0.795313               39544
6                               0.796727               23037
  bivalent_dose_recip_count eligible_recipient_count redacted
1                     11203                    51780       No
2                      2104                    19274       No
3                       139                      759       No
4                      4223                    34657       No
5                     10069                    59905       No
6                      6981                    36576       No
```

Q13. What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2022-11-15"?

```
mean(sd.nov15$percent_of_population_fully_vaccinated, na.rm=TRUE) # na values are removed
```

```
[1] 0.7369099
```

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 "2022-11-15"?

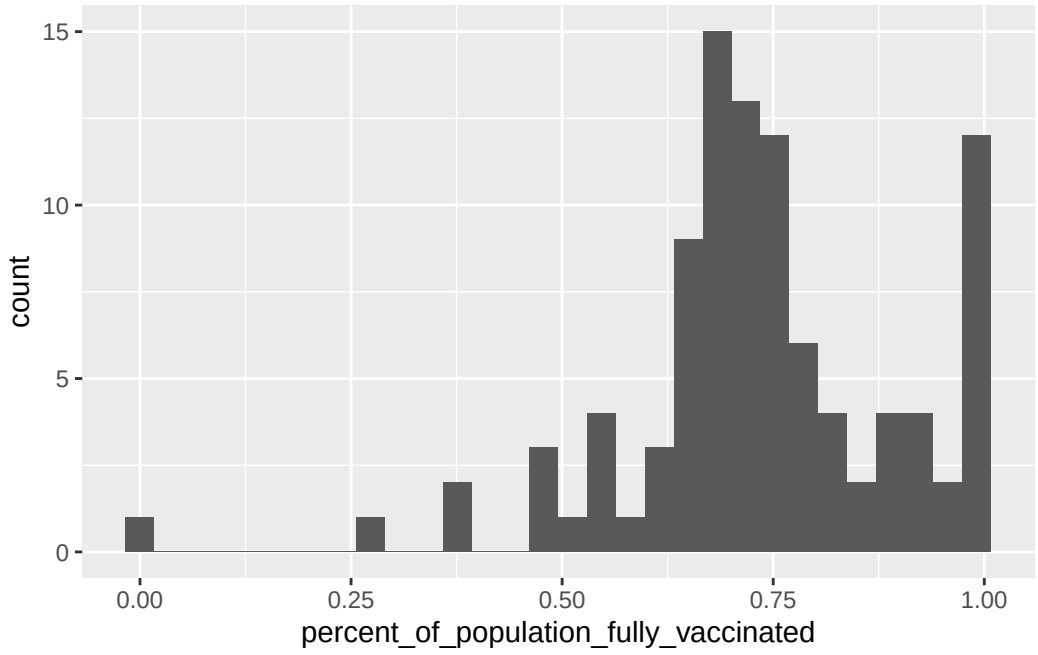
```
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.2.2
```

```
ggplot(sd.nov15) +
  geom_histogram(aes(x = percent_of_population_fully_vaccinated))
```

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

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



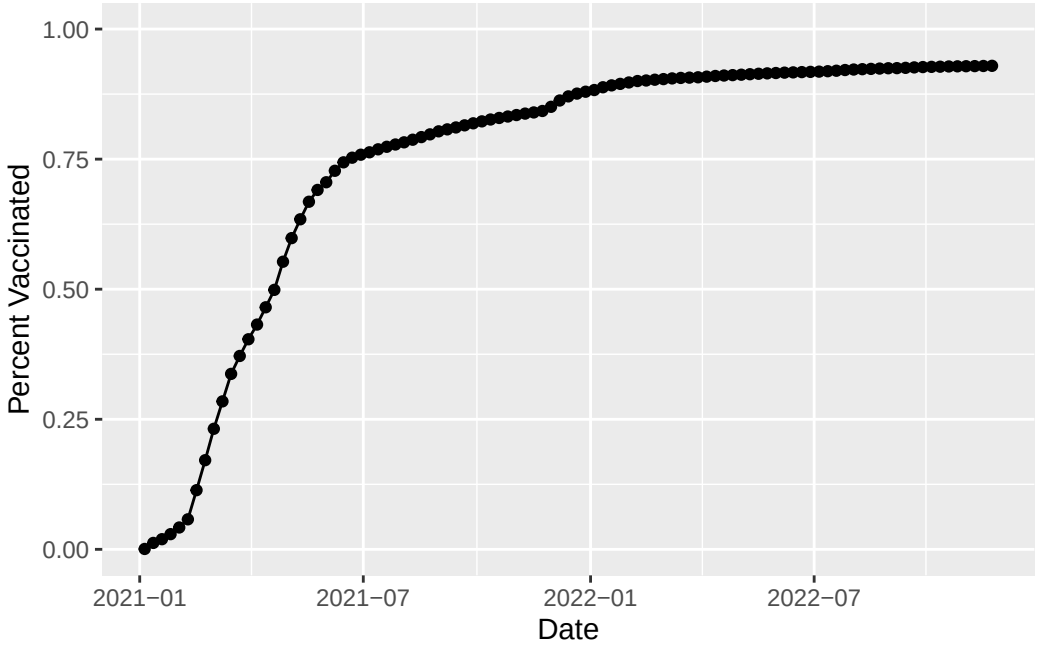
La Jolla/UCSD

```
# isolate La Jolla data and print population age 5+ on first day of data
ucsd <- filter(sd, zip_code_tabulation_area=="92037")
ucsd[1,]$age5_plus_population
```

```
[1] 36144
```

```
# time rate course for vaccination at UCSD

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




```
# 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)
```

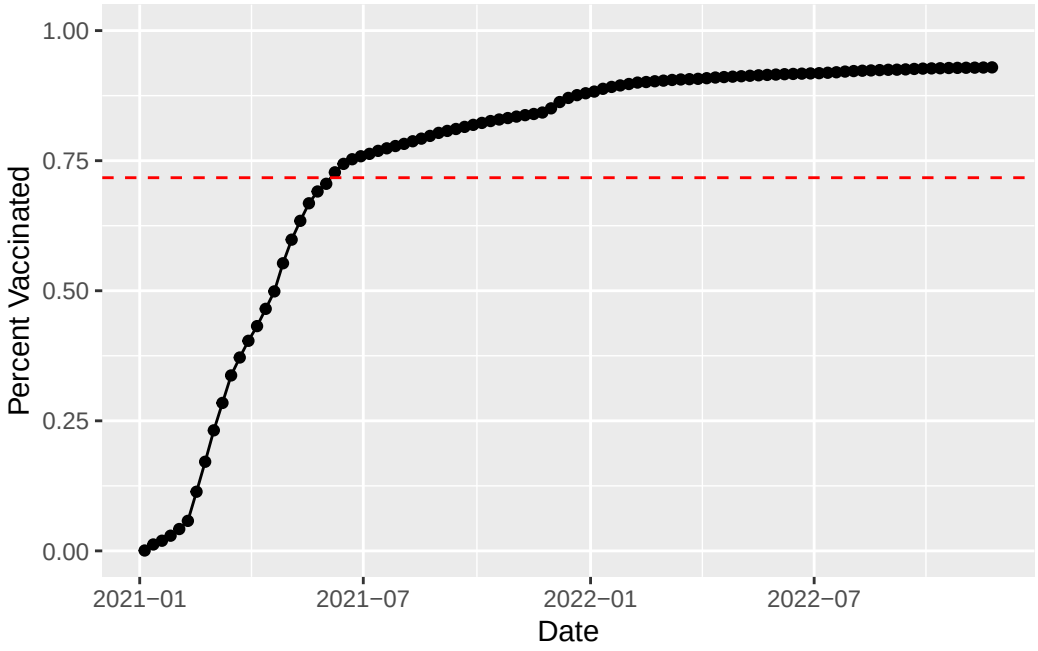
```
  as_of_date zip_code_tabulation_area local_health_jurisdiction        county
1 2022-11-15                   92236                 Riverside     Riverside
2 2022-11-15                   92130                 San Diego     San Diego
3 2022-11-15                   94121           San Francisco San Francisco
4 2022-11-15                   94551                 Alameda       Alameda
5 2022-11-15                   94112           San Francisco San Francisco
6 2022-11-15                   94303             Santa Clara   Santa Clara
  vaccine_equity_metric_quartile                vem_source
1                              1 Healthy Places Index Score
2                              4 Healthy Places Index Score
3                              4 Healthy Places Index Score
4                              4 Healthy Places Index Score
5                              3 Healthy Places Index Score
6                              3 Healthy Places Index Score
  age12_plus_population age5_plus_population tot_population
1               38505.3                42923          45477
2               46300.3                53102          56134
3               39105.0                41363          43616
4               38947.9                43399          47227
5               75681.8                81107          84707
6               40033.3                44989          48244
  persons_fully_vaccinated persons_partially_vaccinated
1                    30465                         3858
2                    52380                         5751
3                    36566                         2373
4                    32557                         2333
5                    78358                         4646
6                    41275                         4175
  percent_of_population_fully_vaccinated
1                               0.669899
2                               0.933124
3                               0.838362
4                               0.689373
5                               0.925048
6                               0.855547
  percent_of_population_partially_vaccinated
1                                   0.084834
2                                   0.102451
```

```
3                                    0.054407
4                                    0.049400
5                                    0.054848
6                                    0.086539
  percent_of_population_with_1_plus_dose booster_recip_count
1                               0.754733               12943
2                               1.000000               34821
3                               0.892769               28345
4                               0.738773               20223
5                               0.979896               56744
6                               0.942086               26288
  bivalent_dose_recip_count eligible_recipient_count redacted
1                      1395                    30375       No
2                     11203                    51780       No
3                     10994                    36013       No
4                      5568                    32234       No
5                     16019                    77580       No
6                      8573                    40853       No
```

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 "2022-11-15". Add this as a straight horizontal line to your plot from above with the geom_hline() function?

```
avg.vaccinated <- mean(vax.36$percent_of_population_fully_vaccinated)

ggplot(ucsd) +
  aes(x = as_of_date,
      percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) +
  geom_hline(yintercept=avg.vaccinated,linetype=2,color="red") +
  ylim(c(0,1)) +
  labs(x="Date", y="Percent Vaccinated")
```



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 "2022-11-15"?

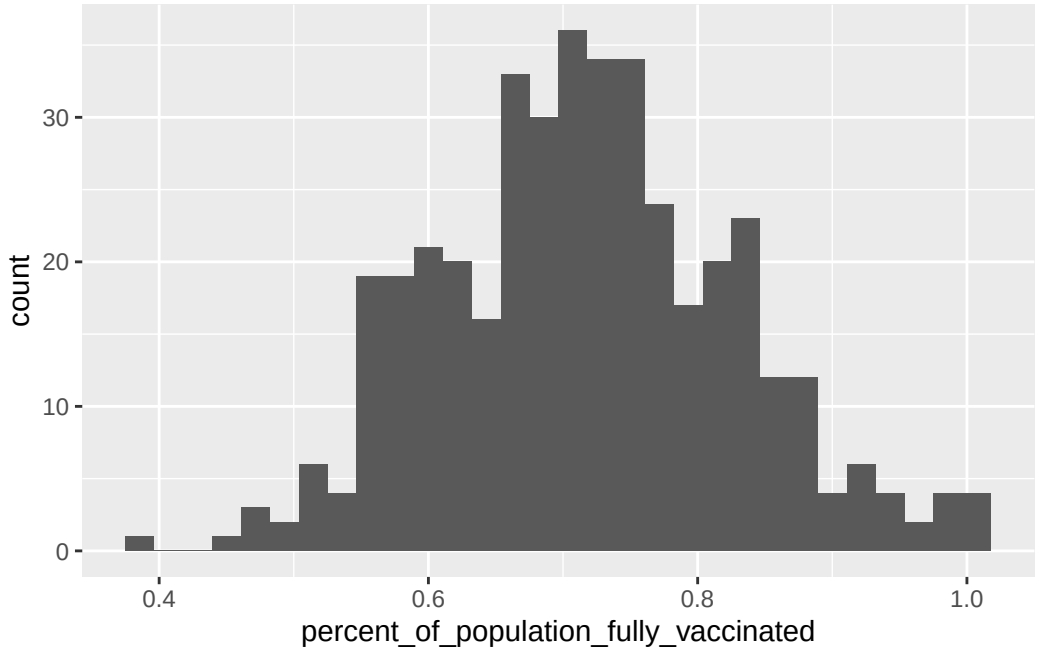
```
summary(vax.36$percent_of_population_fully_vaccinated)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.3785  0.6396  0.7155  0.7173  0.7880  1.0000 
```

Q18. Using ggplot generate a histogram of this data.

```
ggplot(vax.36) +
  geom_histogram(aes(x=percent_of_population_fully_vaccinated))
```

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



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

```
  percent_of_population_fully_vaccinated
1                               0.546646
```

92040 less than calculated above

```
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
```

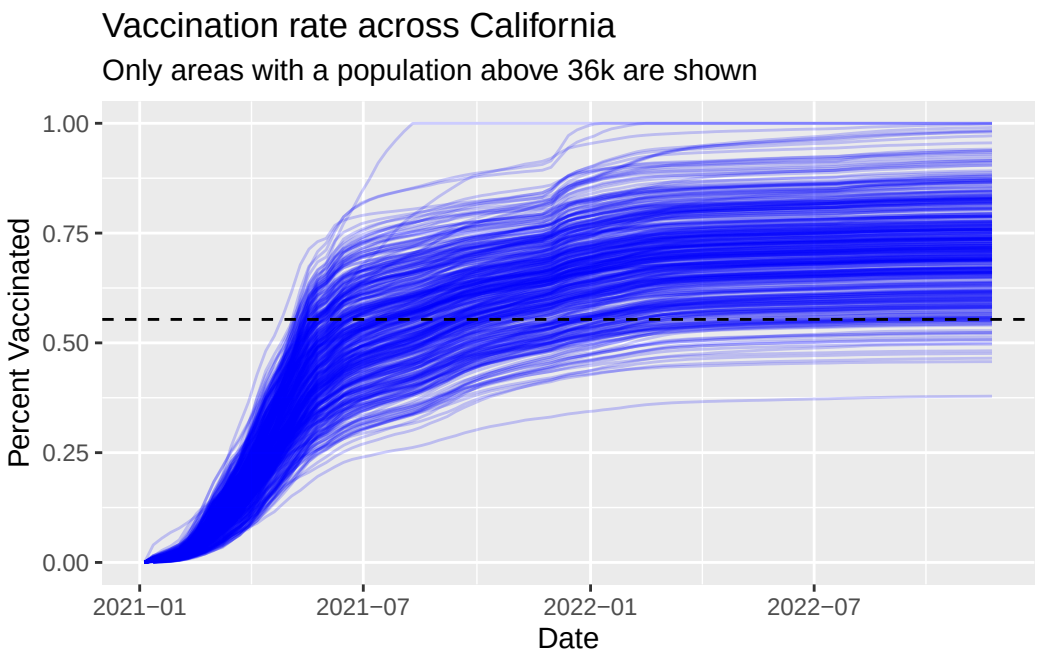
92109 less than calculated above

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


ggplot(vax.36.all) +
  aes(x = as_of_date,
      y = 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 = vax.36.all.mean, linetype=2)
```

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



Q21. How do you feel about coming to class in person after Thanksgiving break?

Honestly probably won't come, but that's more because I prefer to do the labs on my own. I appreciate your asking though!