class16

Victor Yu

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

```
  as_of_date zip_code_tabulation_area local_health_jurisdiction          county
1 2021-01-05                    93609                    Fresno          Fresno
2 2021-01-05                    94086               Santa Clara     Santa Clara
3 2021-01-05                    94304               Santa Clara     Santa Clara
4 2021-01-05                    94110             San Francisco   San Francisco
5 2021-01-05                    93420           San Luis Obispo San Luis Obispo
6 2021-01-05                    93454             Santa Barbara   Santa Barbara
  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                              2 Healthy Places Index Score
  age12_plus_population age5_plus_population tot_population
1                4396.3                 4839           5177
2               42696.0                46412          50477
3                3263.5                 3576           3852
4               64350.7                68320          72380
5               26694.9                29253          30740
6               32043.4                36446          40432
  persons_fully_vaccinated persons_partially_vaccinated
1                       NA                           NA
2                       11                          640
3                       NA                           NA
4                       18                         1262
5                       NA                           NA
6                       NA                           NA
```

```
  percent_of_population_fully_vaccinated
1                                     NA
2                               0.000218
3                                     NA
4                               0.000249
5                                     NA
6                                     NA
  percent_of_population_partially_vaccinated
1                                         NA
2                                   0.012679
3                                         NA
4                                   0.017436
5                                         NA
6                                         NA
  percent_of_population_with_1_plus_dose booster_recip_count
1                                     NA                  NA
2                               0.012897                  NA
3                                     NA                  NA
4                               0.017685                  NA
5                                     NA                  NA
6                                     NA                  NA
  bivalent_dose_recip_count eligible_recipient_count
1                        NA                        1
2                        NA                       11
3                        NA                        6
4                        NA                       18
5                        NA                        4
6                        NA                        5
                                                               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? Answer: Column 10

Q2. What column details the Zip code tabulation area? Answer: Column 2

Q3. What is the earliest date in this dataset? Answer: 2021-01-05

```r
#Use table to look at all the dates and find th earliest one. It's already in order

table(vax["as_of_date"])
```

```
as_of_date
2021-01-05 2021-01-12 2021-01-19 2021-01-26 2021-02-02 2021-02-09 2021-02-16
      1764       1764       1764       1764       1764       1764       1764
2021-02-23 2021-03-02 2021-03-09 2021-03-16 2021-03-23 2021-03-30 2021-04-06
      1764       1764       1764       1764       1764       1764       1764
2021-04-13 2021-04-20 2021-04-27 2021-05-04 2021-05-11 2021-05-18 2021-05-25
      1764       1764       1764       1764       1764       1764       1764
2021-06-01 2021-06-08 2021-06-15 2021-06-22 2021-06-29 2021-07-06 2021-07-13
      1764       1764       1764       1764       1764       1764       1764
2021-07-20 2021-07-27 2021-08-03 2021-08-10 2021-08-17 2021-08-24 2021-08-31
      1764       1764       1764       1764       1764       1764       1764
2021-09-07 2021-09-14 2021-09-21 2021-09-28 2021-10-05 2021-10-12 2021-10-19
      1764       1764       1764       1764       1764       1764       1764
2021-10-26 2021-11-02 2021-11-09 2021-11-16 2021-11-23 2021-11-30 2021-12-07
      1764       1764       1764       1764       1764       1764       1764
2021-12-14 2021-12-21 2021-12-28 2022-01-04 2022-01-11 2022-01-18 2022-01-25
      1764       1764       1764       1764       1764       1764       1764
2022-02-01 2022-02-08 2022-02-15 2022-02-22 2022-03-01 2022-03-08 2022-03-15
      1764       1764       1764       1764       1764       1764       1764
2022-03-22 2022-03-29 2022-04-05 2022-04-12 2022-04-19 2022-04-26 2022-05-03
      1764       1764       1764       1764       1764       1764       1764
2022-05-10 2022-05-17 2022-05-24 2022-05-31 2022-06-07 2022-06-14 2022-06-21
      1764       1764       1764       1764       1764       1764       1764
2022-06-28 2022-07-05 2022-07-12 2022-07-19 2022-07-26 2022-08-02 2022-08-09
      1764       1764       1764       1764       1764       1764       1764
2022-08-16 2022-08-23 2022-08-30 2022-09-06 2022-09-13 2022-09-20 2022-09-27
      1764       1764       1764       1764       1764       1764       1764
2022-10-04 2022-10-11 2022-10-18 2022-10-25 2022-11-01 2022-11-08 2022-11-15
      1764       1764       1764       1764       1764       1764       1764
2022-11-22 2022-11-29 2022-12-06 2022-12-13 2022-12-20 2022-12-27 2023-01-03
      1764       1764       1764       1764       1764       1764       1764
2023-01-10 2023-01-17 2023-01-24 2023-01-31 2023-02-07 2023-02-14 2023-02-21
      1764       1764       1764       1764       1764       1764       1764
2023-02-28 2023-03-07
      1764       1764
```

Q4. What is the latest date in this dataset?

Answer: 2023-03-07

```r
#Table is already in order by dates/ We can just observe the last row
tail(vax)
```

```
       as_of_date zip_code_tabulation_area local_health_jurisdiction
201091 2023-03-07                    93662                    Fresno
201092 2023-03-07                    94801              Contra Costa
201093 2023-03-07                    93668                    Fresno
201094 2023-03-07                    93704                    Fresno
201095 2023-03-07                    94510                    Solano
201096 2023-03-07                    93726                    Fresno
             county vaccine_equity_metric_quartile                 vem_source
201091       Fresno                              1 Healthy Places Index Score
201092 Contra Costa                              1 Healthy Places Index Score
201093       Fresno                              1    CDPH-Derived ZCTA Score
201094       Fresno                              1 Healthy Places Index Score
201095       Solano                              4 Healthy Places Index Score
201096       Fresno                              1 Healthy Places Index Score
       age12_plus_population age5_plus_population tot_population
201091               24501.3                28311          30725
201092               25273.6                29040          31210
201093                1013.4                 1199           1219
201094               24803.5                27701          29740
201095               24819.2                27056          28350
201096               33707.7                39067          42824
       persons_fully_vaccinated persons_partially_vaccinated
201091                    20088                         2150
201092                    27375                         2309
201093                      644                           74
201094                    17887                         1735
201095                    22648                         2264
201096                    24121                         2682
       percent_of_population_fully_vaccinated
201091                               0.653800
201092                               0.877123
201093                               0.528302
201094                               0.601446
201095                               0.798871
201096                               0.563259
       percent_of_population_partially_vaccinated
201091                                    0.069976
```

```
201092                                   0.073983
201093                                   0.060705
201094                                   0.058339
201095                                   0.079859
201096                                   0.062628
       percent_of_population_with_1_plus_dose booster_recip_count
201091                                 0.723776               10072
201092                                 0.951106               14782
201093                                 0.589007                 312
201094                                 0.659785               10435
201095                                 0.878730               16092
201096                                 0.625887               12104
       bivalent_dose_recip_count eligible_recipient_count redacted
201091                      2578                    20066       No
201092                      5342                    27282       No
201093                        66                      644       No
201094                      4154                    17822       No
201095                      8797                    22501       No
201096                      3585                    24062       No
```

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

```
[1] "2023-03-07"
```

```r
#Quick overview of dataset
skimr::skim(vax)
```

Table 1: Data summary

Name	vax
Number of rows	201096
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	114	0
local_health_jurisdiction	0	1	0	15	570	62	0
county	0	1	0	15	570	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	1817.38	90001	92257.75	93658.50	95380.50	97635.0	
vaccine_equity_metric_quartile	9918	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	
age12_plus_population	0	1.00	18895.04	18993.87	0	1346.95	13685.10	31756.12	88556.7	
age5_plus_population	0	1.00	20875.24	21105.97	0	1460.50	15364.00	34877.00	101902.0	
tot_population	9804	0.95	23372.77	22628.50	12	2126.00	18714.00	38168.00	111165.0	
persons_fully_vaccinated	16621	0.92	13990.39	15073.66	11	932.00	8589.00	23346.00	87575.0	
persons_partially_vaccinated	16621	0.92	1702.31	2033.32	11	165.00	1197.00	2536.00	39973.0	
percent_of_population_fully_vaccinated	20965	0.90	0.57	0.25	0	0.42	0.61	0.74	1.0	
percent_of_population_partially_vaccinated	20965	0.90	0.08	0.09	0	0.05	0.06	0.08	1.0	
percent_of_population_with_1_plus_dose	22009	0.89	0.63	0.24	0	0.49	0.67	0.81	1.0	
booster_recip_count	72997	0.64	5882.76	7219.00	11	300.00	2773.00	9510.00	59593.0	
bivalent_dose_recip_count	158776	0.21	2978.23	3633.03	11	193.00	1467.50	4730.25	27694.0	
eligible_recipient_count	0	1.00	12830.83	14928.64	0	507.00	6369.00	22014.00	87248.0	

```
#vax$persons_fully_vaccinated
```

Q5. How many numeric columns are in this dataset? Answer: 13 columns

Q6: Note that there are "missing values" in the dataset. How many NA values there in the persons_fully_vaccinated column? Answer: 16621 NA vales

```
#Use 'is.na' to give a T/F matrix
#table () it to give you the count of each

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

```

 FALSE   TRUE 
184475  16621 
```

```r
#sum () adds up the nukmber of TRUE. We can store this into n.missing to use it
n.missing <- sum (is.na(vax$persons_fully_vaccinated))
round ((n.missing / nrow(vax))*100, 2)
```

```
[1] 8.27
```

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

8.27

WORKING WITH DATES

```r
library (lubridate)
```

```

Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

```r
today()
```

```
[1] "2023-03-10"
```

```r
#Specify that we are using the year-month-day format
#For funsies. This will give an Error! today() - vax$as_of_date[1] need vax$as_of_date fir
vax$as_of_date <- ymd (vax$as_of_date)
```

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

```
Time difference of 794 days
```

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

```
Time difference of 791 days
```

```r
#Find the most recent date in the data set
today() - ymd (vax$as_of_date[nrow(vax)])
```

```
Time difference of 3 days
```

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

Answers: 3 days (as of 3/10/2023)

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

Answer: 114 unique dates

```r
nrow(table(vax$as_of_date))
```

```
[1] 114
```

Zip Codes

```r
#installed zipcodeR package
library("zipcodeR")
```

```r
#geocode_zip gives certain zip
geocode_zip('92037')
```

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

```r
#Inputting 2 zip codes with zip_distance gives you the distance between them (IN MILES)
zip_distance('92037', '92109')
```

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

```r
# reverse_zipcode pulls out all the related information tied to the zip code
# we can store this in zip_data
zip_data <- reverse_zipcode(c('92037', "92109"))
```

```r
#Method 1: Subset SD county
sd <- vax$county == "San Diego"
sdx <- vax[sd,]
```

```r
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
#Method 2

sd.2 <- filter(vax, county == "San Diego")
nrow(sd.2)
```

```
[1] 12198
```

```r
#Keep in mind: sd2 & sdx both "San Diego-sorted" dataframes
nrow(table(sdx$zip_code_tabulation_area))
```

```
[1] 107
```

Q11. How many distinct zip codes are listed for San Diego County? Answer: 107 zip codes in SD

```
#Find index of row largest tot_population
#Use the index to find the zip code is matches with

high <- which.max(sdx$age12_plus_population)
sdx[high, "zip_code_tabulation_area"]
```

```
[1] 92154
```

Q12. What San Diego County Zip code are ahas the largest 12+ population in this dataset? Answer:92154 largest population in dataset

```
#Using dplyr to filter the df
sd.date <- filter (vax, county == "San Diego" & as_of_date == "2023-02-28")
```

```
#Remove NA row first & new df for ease

sd.ppfv <- sd.date[is.na(sd.date$percent_of_population_fully_vaccinated) == 0,]

#take average new dataframe without NA rows

mean(sd.ppfv$percent_of_population_fully_vaccinated)
```

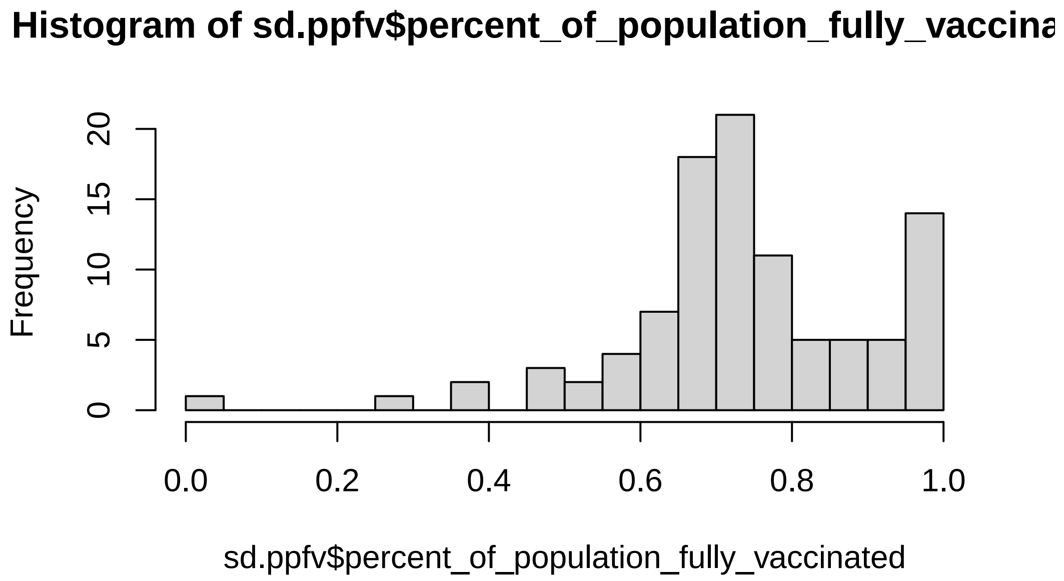
```
[1] 0.7401687
```

Q13. What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2023-02-28"

Answer: 0.7401687

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"?

```
#Base R plots
hist(sd.ppfv$percent_of_population_fully_vaccinated, breaks=20)
```



UCSD & La Jolla

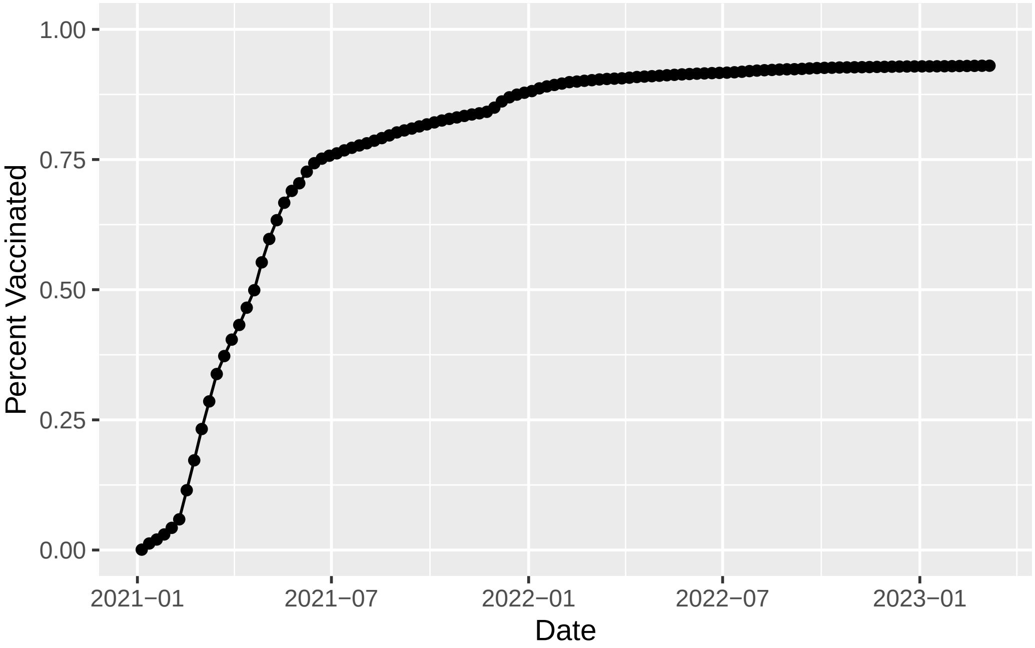
```
#dplyr filter by area code
#T/F dataframe & sdx acutal dataframe that's sorted

ucsd <- filter(sdx, 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:

```
library (ggplot2)
```

```
# Fill in the ggplot code from the lab manual
ucsdplot <- 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")
ucsdplot
```



```
# Subset to all CA areas with a population as large as 92037
vax.36 <- filter(vax, age5_plus_population > 36144 &
                as_of_date == "2023-02-28")

head(vax.36)
```

```
  as_of_date zip_code_tabulation_area local_health_jurisdiction         county
1 2023-02-28                    91710            San Bernardino San Bernardino
2 2023-02-28                    92231                  Imperial       Imperial
3 2023-02-28                    93436             Santa Barbara  Santa Barbara
4 2023-02-28                    95037               Santa Clara    Santa Clara
5 2023-02-28                    92234                 Riverside      Riverside
6 2023-02-28                    95120               Santa Clara    Santa Clara
  vaccine_equity_metric_quartile                 vem_source
1                              3 Healthy Places Index Score
2                              1 Healthy Places Index Score
```

```
3                               2 Healthy Places Index Score
4                               4 Healthy Places Index Score
5                               1 Healthy Places Index Score
6                               4 Healthy Places Index Score
  age12_plus_population age5_plus_population tot_population
1               79765.1                86612          91773
2               32448.6                36867          40064
3               46236.9                52318          56323
4               43786.2                48583          51994
5               46401.1                51202          54357
6               32743.9                36636          38122
  persons_fully_vaccinated persons_partially_vaccinated
1                    53009                         4698
2                    71106                        39909
3                    34961                         4161
4                    43309                         2824
5                    38397                         4954
6                    35627                         2201
  percent_of_population_fully_vaccinated
1                               0.577610
2                               1.000000
3                               0.620723
4                               0.832961
5                               0.706386
6                               0.934552
  percent_of_population_partially_vaccinated
1                                   0.051192
2                                   0.996131
3                                   0.073877
4                                   0.054314
5                                   0.091138
6                                   0.057736
  percent_of_population_with_1_plus_dose booster_recip_count
1                               0.628802               30093
2                               1.000000               29254
3                               0.694600               19444
4                               0.887275               29756
5                               0.797524               21318
6                               0.992288               28307
  bivalent_dose_recip_count eligible_recipient_count redacted
1                     10464                    52875       No
2                      5301                    70768       No
3                      7056                    34857       No
```

```
4                       12364                    43137       No
5                        7771                    38367       No
6                       14895                    35476       No
```

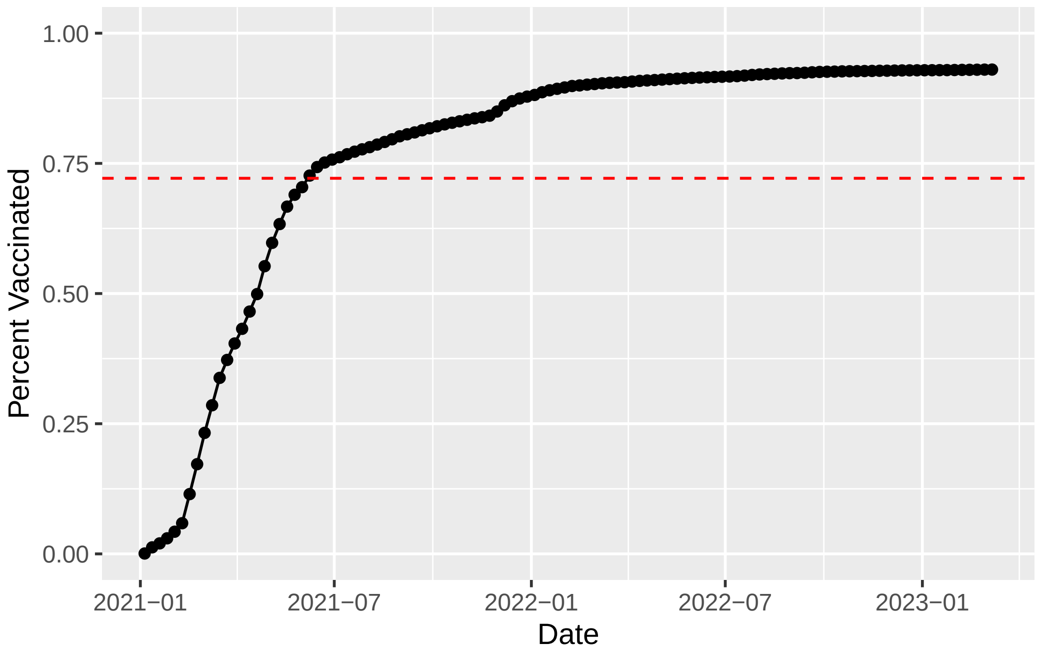
Q16. Calculate the mean “Percent of Population Fully Vaccinated” for ZIP code areas with a population as large as 92037 (La Jolla) as_of_date “2023-02-28”. Add this as a straight horizontal line to your plot from above with the geom_hline() function?

Anwer: 0.72149

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

```
[1] 0.7213907
```

```r
#Adding existing ucsd.plot with geom_hline()
ucsdplot + geom_hline(yintercept = mean(vax.36$percent_of_population_fully_vaccinated), co
```



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 “2023-02-28”?

```
#Fivenum() fucntions works on data., Releases a output of min, lowerQ, median, upperQ, max
fivenum(vax.36$percent_of_population_fully_vaccinated)
```

```
[1] 0.3804340 0.6458120 0.7181270 0.7907105 1.0000000
```

```
#Mean
vax36.mean <- mean(vax.36$percent_of_population_fully_vaccinated)

sixnum <- c(fivenum(vax.36$percent_of_population_fully_vaccinated), vax36.mean)
```

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

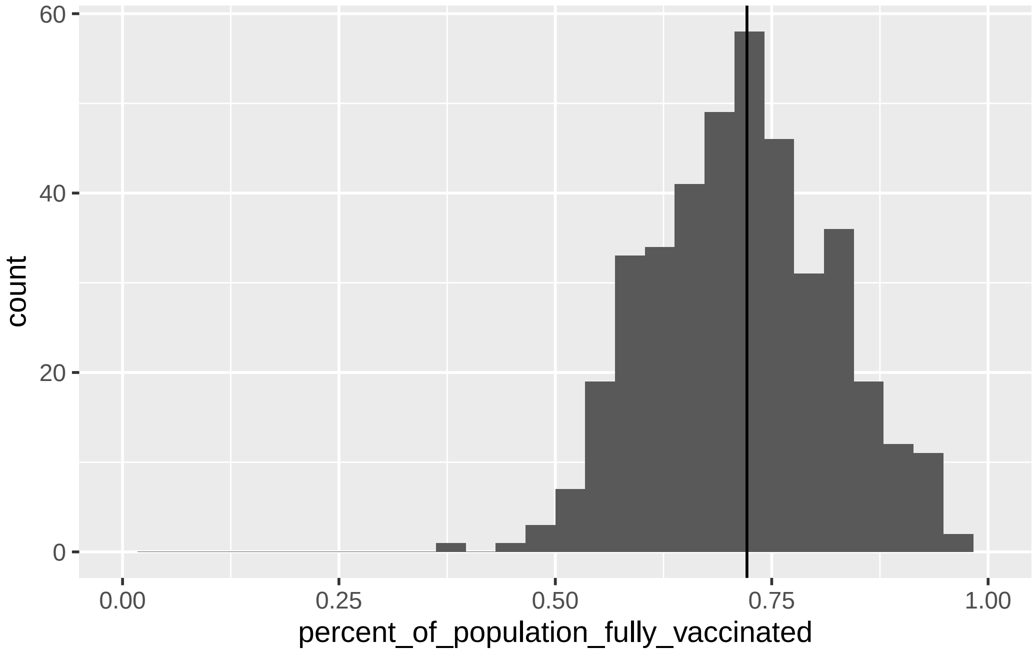
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.3804  0.6458  0.7181  0.7214  0.7907  1.0000 
```

Q18. Using ggplot generate a histogram of this data.

```
vax36.plot <- ggplot(vax.36) +
  aes(x=percent_of_population_fully_vaccinated) +
  xlim(0,1) +
  geom_histogram() + geom_vline(aes(xintercept=vax36.mean))
vax36.plot
```

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

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



```r
#Average for 92040

vax %>% filter(as_of_date == "2023-02-28") %>%
  filter (zip_code_tabulation_area == "92040") %>%
  select(percent_of_population_fully_vaccinated)
```

```
  percent_of_population_fully_vaccinated
1                               0.550469
```

```r
#Average for 92109

vax %>% filter(as_of_date == "2023-02-28") %>%
  filter (zip_code_tabulation_area == "92109") %>%
  select(percent_of_population_fully_vaccinated)
```

```
  percent_of_population_fully_vaccinated
1                                0.69453
```

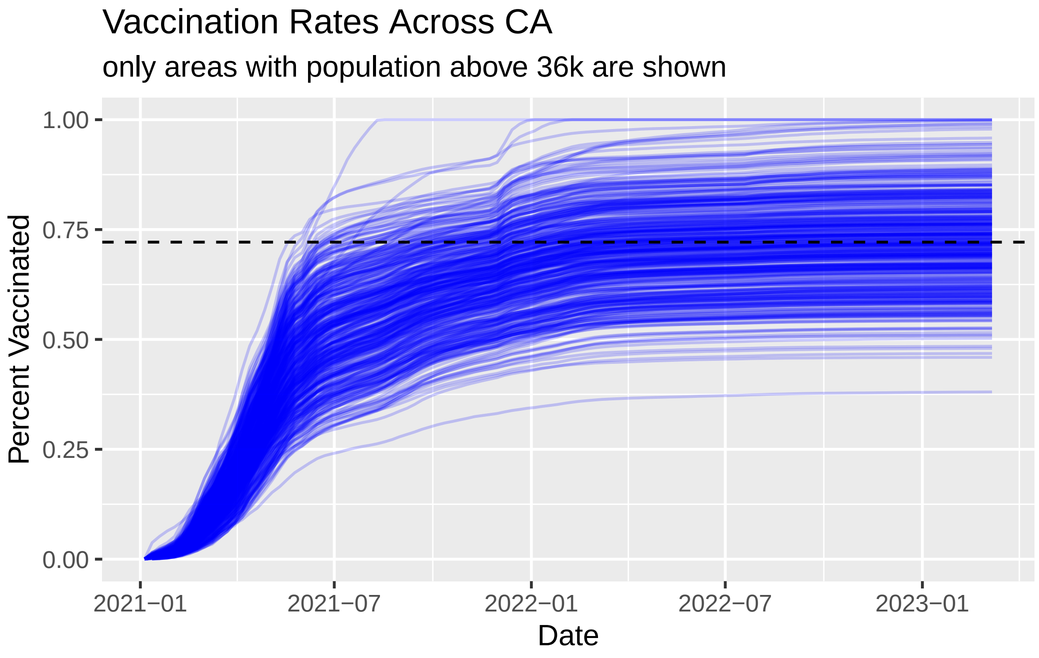
Q19. Is the 92109 and 92040 ZIP code areas above or below the average value you calculated for all these above?

Answer: The two averages are .550469, .69453 which are both below average

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

```r
vax36.all <- filter(vax, age5_plus_population > 36144)
ggplot(vax36.all) +
 aes(as_of_date,
 percent_of_population_fully_vaccinated,
 group=zip_code_tabulation_area) +
 geom_line(alpha=0.2, color="blue") +
 ylim(0,1) +
 labs(x="Date", y="Percent Vaccinated",
 title="Vaccination Rates Across CA",
 subtitle="only areas with population above 36k are shown") +
 geom_hline(yintercept = vax36.mean, linetype="dashed")
```

```
Warning: Removed 183 rows containing missing values (`geom_line()`).
```


Vaccination Rates Across CA
only areas with population above 36k are shown
Percent Vaccinated
Date