Class 14: Vaccination Rate Mini-Project

Taylor F. (A59010460)

3/4/2022

#Intro to Vax Data

```
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
```

```
tail(vax)
```

```
##        as_of_date zip_code_tabulation_area local_health_jurisdiction
## 107599 2022-03-01                   91945                San Diego
## 107600 2022-03-01                   91741              Los Angeles
## 107601 2022-03-01                   91768              Los Angeles
## 107602 2022-03-01                   91345              Los Angeles
## 107603 2022-03-01                   91356              Los Angeles
## 107604 2022-03-01                   94402                San Mateo
##             county vaccine_equity_metric_quartile                vem_source
## 107599   San Diego                             2 Healthy Places Index Score
## 107600 Los Angeles                             3 Healthy Places Index Score
## 107601 Los Angeles                             1 Healthy Places Index Score
## 107602 Los Angeles                             2 Healthy Places Index Score
## 107603 Los Angeles                             3 Healthy Places Index Score
## 107604   San Mateo                             4 Healthy Places Index Score
##        age12_plus_population age5_plus_population persons_fully_vaccinated
## 107599               22820.5                25486                    18164
## 107600               22895.7                25243                    19051
## 107601               29837.1                32658                    20587
## 107602               16767.4                18029                    14872
## 107603               26392.1                28379                    22863
## 107604               21862.1                24150                    23094
##        persons_partially_vaccinated percent_of_population_fully_vaccinated
## 107599                         4032                                0.712705
## 107600                         1438                                0.754704
## 107601                         2467                                0.630382
## 107602                         1371                                0.824893
## 107603                         2114                                0.805631
## 107604                         1697                                0.956273
##        percent_of_population_partially_vaccinated
## 107599                                   0.158205
## 107600                                   0.056966
## 107601                                   0.075540
## 107602                                   0.076044
## 107603                                   0.074492
## 107604                                   0.070269
##        percent_of_population_with_1_plus_dose booster_recip_count redacted
## 107599                               0.870910               6542       No
## 107600                               0.811670              10331       No
```

```
## 107601                          0.705922               8694       No
## 107602                          0.900937               6715       No
## 107603                          0.880123              12372       No
## 107604                          1.000000              16049       No
```

```
#We can also us skimr to find out which date is the latest
```

Question 1: What column details the total number of people fully vaccinated?

The column is called "persons_fully_vaccinated" and is column number 9

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

The column is called "zip_code_tabulation_area" and is column number 2

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

The earliest date is 2021-01-05

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

The latest date is 2022-03-01

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

Table 1: Data summary

Name	vax
Number of rows	107604
Number of columns	15
Column type frequency:	
character	5
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	1307	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	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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	

```
skim_grab <- skimr::skim(vax)
```

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

9

Question 6: Note that there are "missing values" in the dataset. How many NA values there in the persons_fully_vaccinated column?

```
vax_na <- is.na(vax$persons_fully_vaccinated == TRUE)

sum(vax_na)
```

```
## [1] 18338
```

The number of missing values is 18338

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

The percent of of values that are 'N/A' is 17%

Question 8: [Optional]: Why might this data be missing?

Working with Dates

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
today() - ymd("1995-10-13")
```

```
## Time difference of 9641 days
```

```
time_length(today() - ymd("1995-10-13"), "years")
```

```
## [1] 26.39562
```

The as_of_date column of our data is currently not that usable. For example we can't easily do math with it like answering the simple question how many days have passed since data was first recorded:

```
#This will give an Error!
#today() - vax$as_of_date[1]
```

However if we convert our date data into a lubridate format things like this will be much easier as well as plotting time series data later on.

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

#We can now do math using dates in our dataset
today() - vax$as_of_date[1]
```

```
## Time difference of 425 days
```

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

```
## Time difference of 5 days
```

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

There have been 5 days since the last update

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

```
skim_grab[1,8]
```

```
## # A tibble: 1 x 1
##   character.n_unique
##                <int>
## 1                 61
```

Using dplyr

Standard Approach

```
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
```

dplyr Approach

```
vax %>% filter(county == "San Diego") -> myans
```

```
#Filtering out areas with fewer than 10,000 residents
sd.10 <- filter(vax, county == "San Diego" &
                age5_plus_population > 10000)
```

Question 11: How many unique zip codes are in San Diego county

```
head(sd)
```

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 1 2021-01-05                   92130                San Diego San Diego
## 2 2021-01-05                   91945                San Diego San Diego
## 3 2021-01-05                   91917                San Diego San Diego
## 4 2021-01-05                   92103                San Diego San Diego
## 5 2021-01-05                   92075                San Diego San Diego
## 6 2021-01-05                   92084                San Diego San Diego
##   vaccine_equity_metric_quartile                vem_source
## 1                              4 Healthy Places Index Score
## 2                              2 Healthy Places Index Score
## 3                              1    CDPH-Derived ZCTA Score
## 4                              4 Healthy Places Index Score
## 5                              4 Healthy Places Index Score
## 6                              2 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1               46300.3                53102                       61
## 2               22820.5                25486                       NA
## 3                 826.1                  939                       NA
## 4               32146.4                33213                       45
## 5               11136.3                12177                       NA
## 6               42677.7                47784                       12
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                           27                                0.001149
## 2                           NA                                      NA
## 3                           NA                                      NA
## 4                           30                                0.001355
## 5                           NA                                      NA
## 6                           17                                0.000251
##   percent_of_population_partially_vaccinated
## 1                                   0.000508
```

```
## 2                                       NA
## 3                                       NA
## 4                                 0.000903
## 5                                       NA
## 6                                 0.000356
##   percent_of_population_with_1_plus_dose booster_recip_count
## 1                               0.001657                 NA
## 2                                     NA                 NA
## 3                                     NA                 NA
## 4                               0.002258                 NA
## 5                                     NA                 NA
## 6                               0.000607                 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
```

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

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

In some cases, dplyr is not the most effective way to accomplish a task

```
vax %>% filter(county == "San Diego") %>% select(zip_code_tabulation_area) %>% unique() %>% nrow()
```

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

Question 12: Which zip code in San Diego county has the largest 12 + population

```
inds <- order(sd$age12_plus_population, decreasing = TRUE)

#Let's use dplyr
head(arrange(sd, age12_plus_population) %>% select(zip_code_tabulation_area), 1)
```

```
##   zip_code_tabulation_area
## 1                    92132
```

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

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

Question 16:

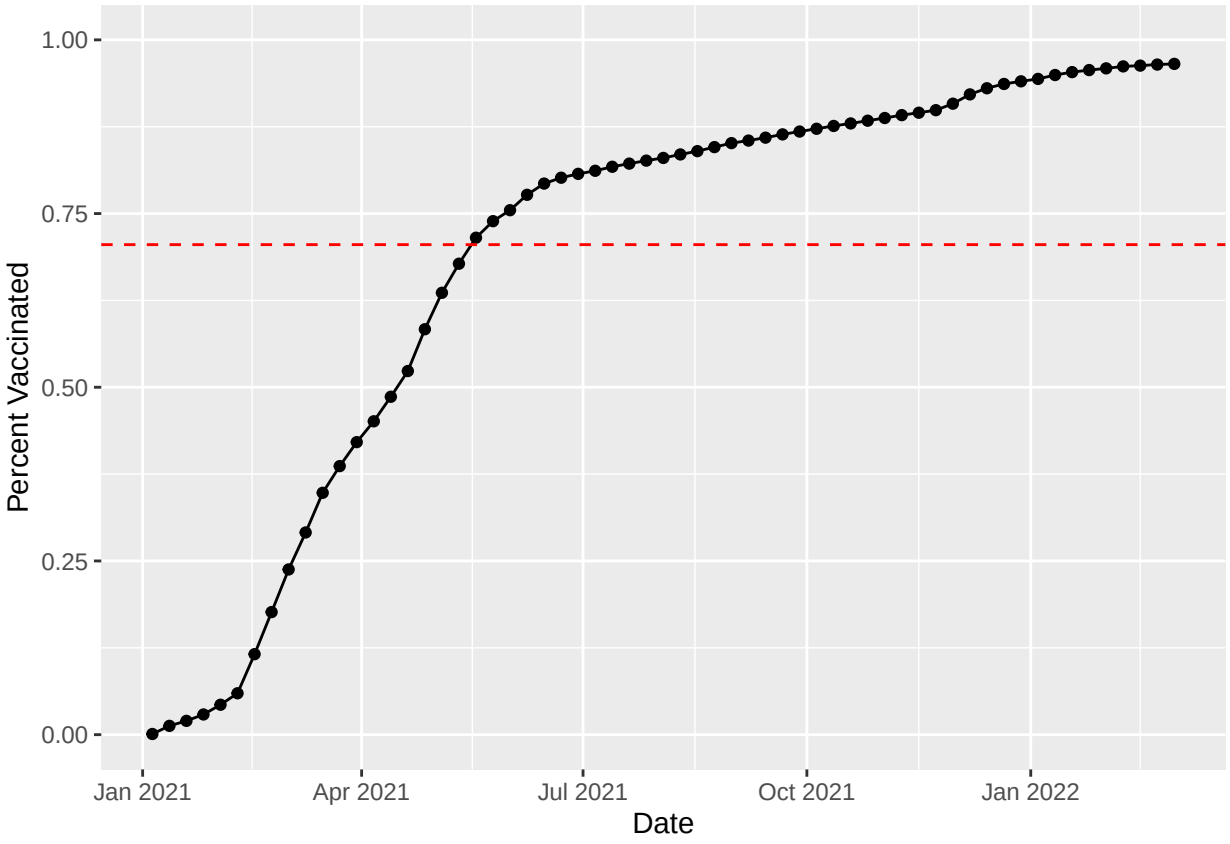
```
sd.now <- filter(sd, as_of_date == "2022-03-01")
avg_vaxx <- mean(sd.now$percent_of_population_fully_vaccinated, na.rm = TRUE)
```

Question 13: What is the overall average “Percent of Population Fully Vaccinated” value for all San Diego “County” as of “2022-02-22”?

The average is 0.7052904

Question 15:

```
library(ggplot2)
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") + geom_hline(yintercept = avg_vaxx, linetype = "dashed", col
```



Question 17:

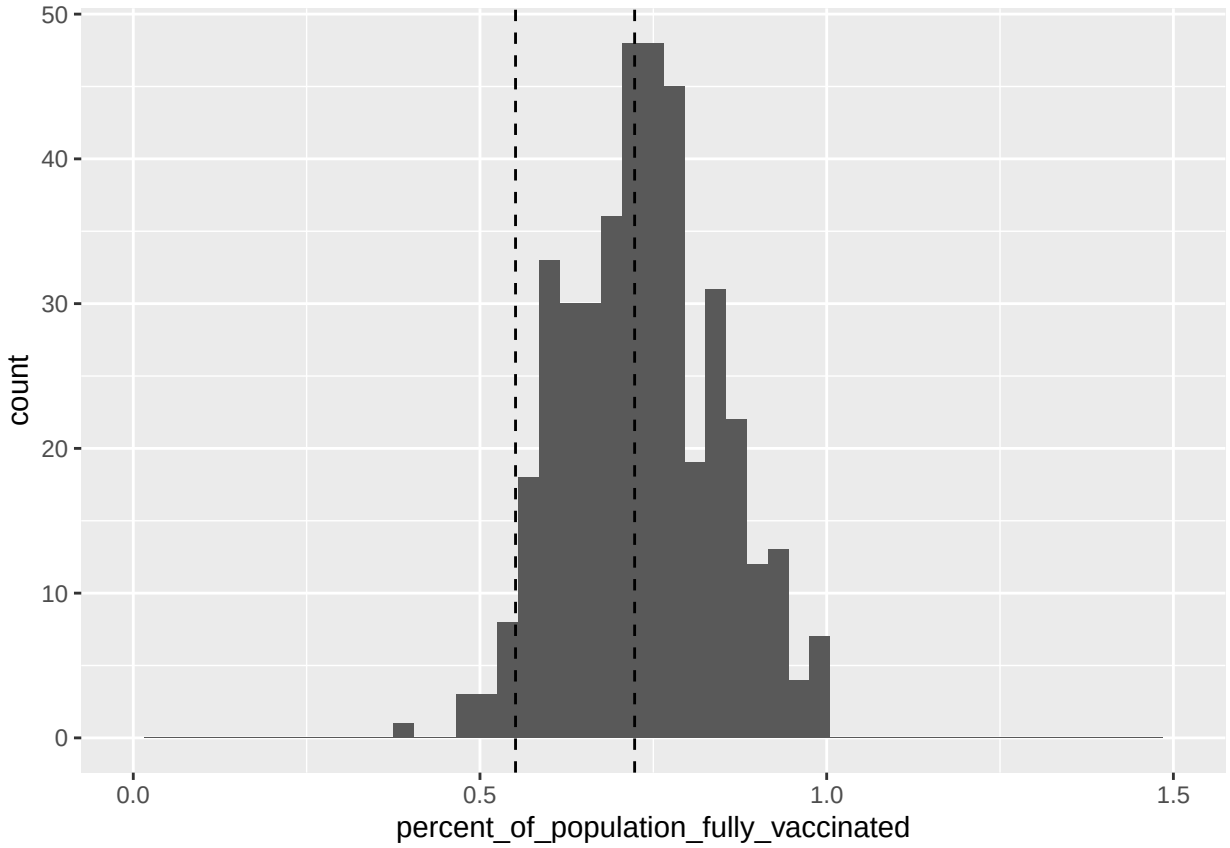
```
vax.36 <- filter(vax, age5_plus_population > 36144 &
                 as_of_date == "2022-02-22")
summary(vax.36$persons_fully_vaccinated)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15406   30551   35305   38118   43420   77457
```

Question 18:

```
ggplot(vax.36, aes(percent_of_population_fully_vaccinated)) + geom_histogram(binwidth = 0.03) + xlim(0,
```

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

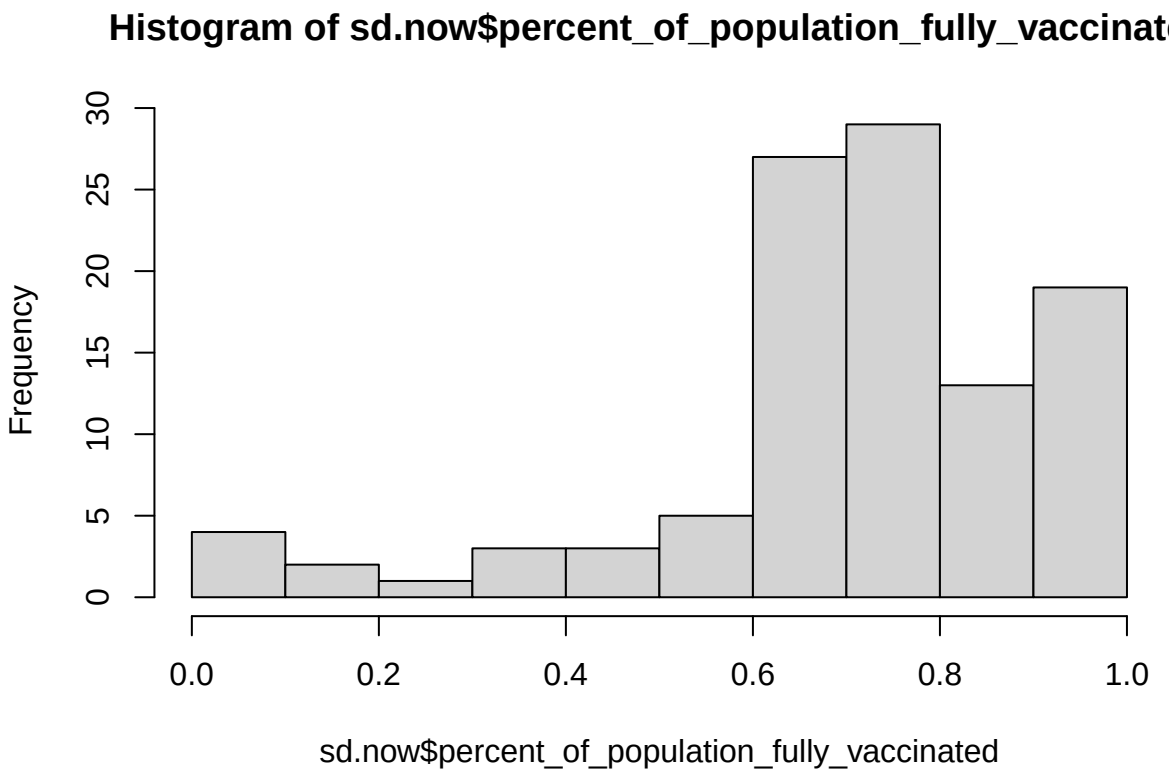


Question 19: Is the 92109 and 92040 ZIP code areas above or below the average value you calculated for all these above?

92109 is above the average, while 92040 is below

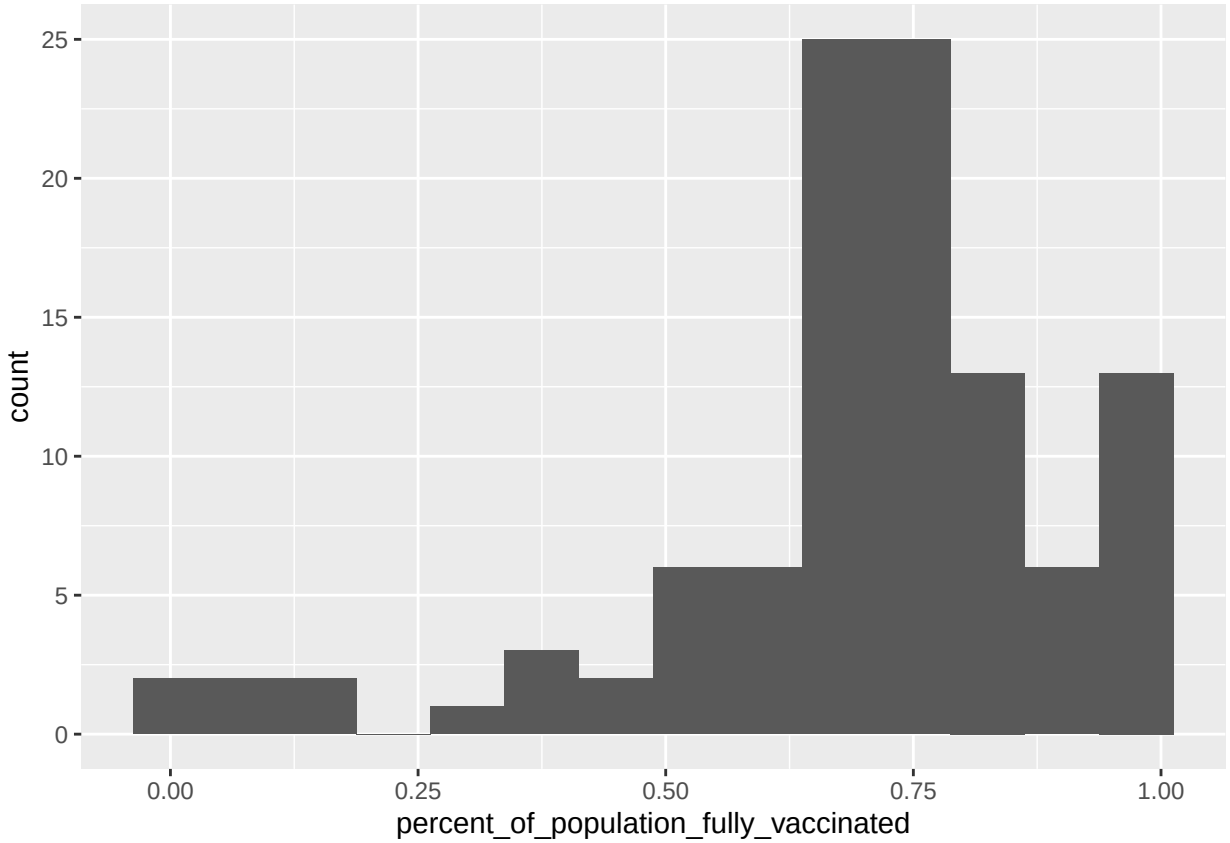
Question 14:

```
hist(sd.now$percent_of_population_fully_vaccinated)
```




```
ggplot(sd.now, aes(percent_of_population_fully_vaccinated)) + geom_histogram(binwidth = 0.075)
```

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

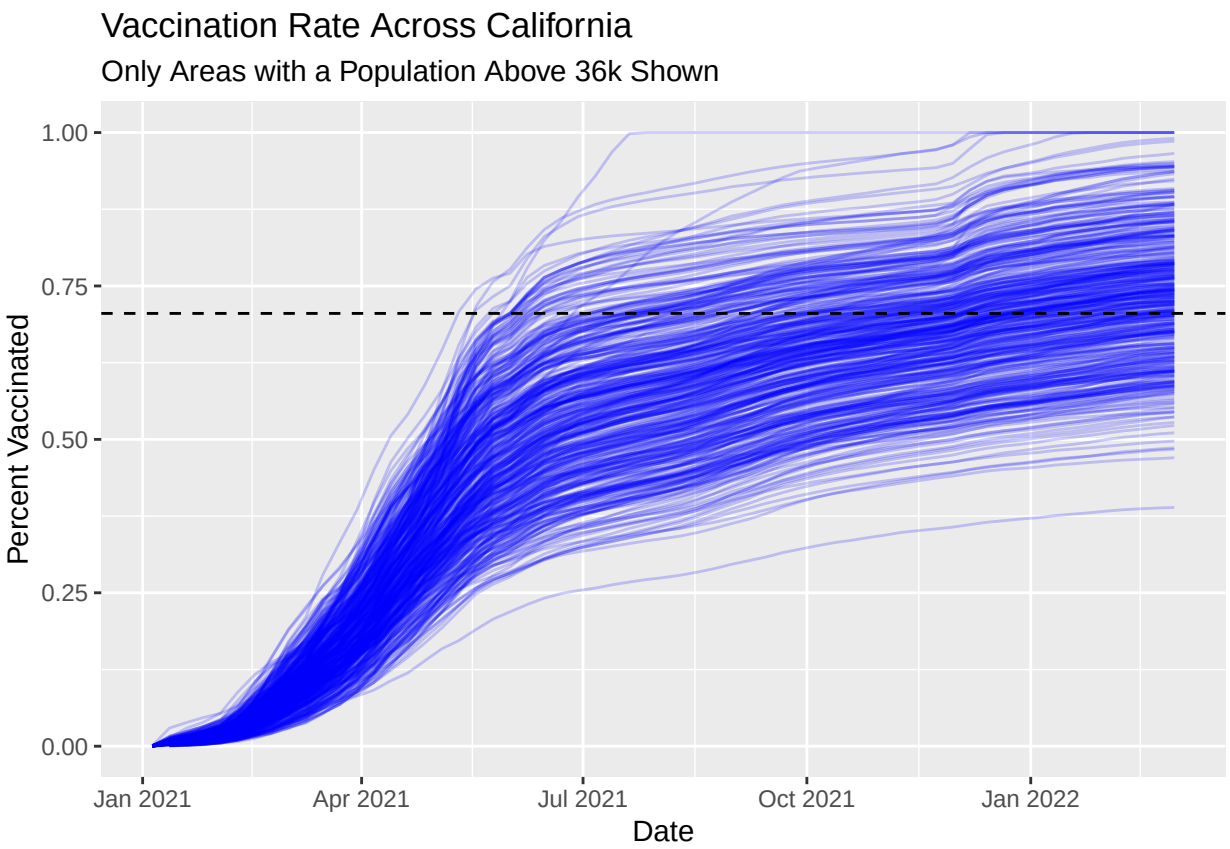


Question 20:

```
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(0,1) +
  labs(x="Date", y="Percent Vaccinated",
       title="Vaccination Rate Across California",
       subtitle="Only Areas with a Population Above 36k Shown") +
  geom_hline(yintercept = avg_vaxx, linetype="dashed")
```

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


Vaccination Rate Across California
Only Areas with a Population Above 36k Shown
Percent Vaccinated
1.00
0.75
0.50
0.25
0.00
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
Date