Class14: Vaccination rate mini project

Sam Altshuler (PID: A59010373)

3/4/2022

Vaccination Rate Mini Project woot woot!

Read in the CA Vaccination Data

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

The total number of people fully vaccinated is found in the `persons_fully_vaccinated` column and the Zipcode tabulation area is found in the `zip_code_tabulation_area` column. The earliest date in this data set is 2021-01-05 (January 5th, 2021) while the latest date in the data set is 2022-03-01 (March 1st, 2022) (this past Tuesday).

```
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	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	

There are 9 columns that are numeric (the zip code column is recognized as a numeric column but isn't really since it's a label) and 5 columns that are character values. In `persons_fully_vaccinated`, there are 18338 values missing (NA) which means that 17% missing (1- `complete_rate` value) (data taken from the readout from the skim function). This data might be missing due to some zip codes might not report their data to the state and people who get shots in some zip codes might not live there.

```
# Way to get the data listed above not from the skim data output
na_val <- sum( is.na(vax$persons_fully_vaccinated) ) / length(vax$persons_fully_vaccinated)
per_missing <- signif(sum( is.na(vax$persons_fully_vaccinated) ) / length(vax$persons_fully_vaccinated)
```

Number of NA values = `r na_val`, percent missing data = 17%.

Play with the dates

Load in the lubridate package to deal with dates easier.

```
library(lubridate)
```

How old am I?

```
today() - ymd("1999-08-17")
```

```
## Time difference of 8235 days
```

I am 8235 days old!

```
time_length(today() - ymd("1999-08-17"), "years")
```

```
## [1] 22.5462
```

I'm 22.5462012 years old!

Now how many days does the vaccination dataset cover?

```
# Convert as_of_date column to the output of the ymd function
vax$as_of_date <- ymd(vax$as_of_date)
vax$as_of_date[nrow(vax)] - vax$as_of_date[1]
```

```
## Time difference of 420 days
```

It’s been 420 days since the state first started recording vaccination data (as of the last day they updated the dataset).

Overall it’s been 423 days since the first date in the dataset.

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

```
## Time difference of 423 days
```

And it’s been 3 days since the last date in the dataset.

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

```
## Time difference of 3 days
```

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

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

Overall, there are 61 unique dates in the dataset.

Working with ZIP codes (focusing on SD area)

There are two ways to approach this: base R or dplyr.

```r
#Base R
# sd < - vax[vax$county == "San Diego", ]
```

```r
#Dplyr and ggplot2 are inside tidyverse
library(tidyverse)
```

```r
sd <- vax %>%
  filter(county == "San Diego")
head(sd, 3)
```

```
##   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
##   vaccine_equity_metric_quartile                 vem_source
## 1                              4 Healthy Places Index Score
## 2                              2 Healthy Places Index Score
## 3                              1    CDPH-Derived ZCTA 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
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                           27                               0.001149
## 2                           NA                                     NA
## 3                           NA                                     NA
##   percent_of_population_partially_vaccinated
## 1                                   0.000508
## 2                                         NA
## 3                                         NA
##   percent_of_population_with_1_plus_dose booster_recip_count
## 1                               0.001657                  NA
## 2                                     NA                  NA
## 3                                     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
```

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

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

There are 107 unique ZIP codes in SD county (according to this dataset).

Which ZIP code has the largest 12+ population?

```
# Use which.max() to find the index of the maximum value in the 12+ population column
sd$zip_code_tabulation_area[which.max(sd$age12_plus_population)]
```

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

```
#using  dplyr
arrange(sd, -age12_plus_population)$zip_code_tabulation_area[1]
```

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

The 92154 ZIP code has the largest population of 12+ individuals.

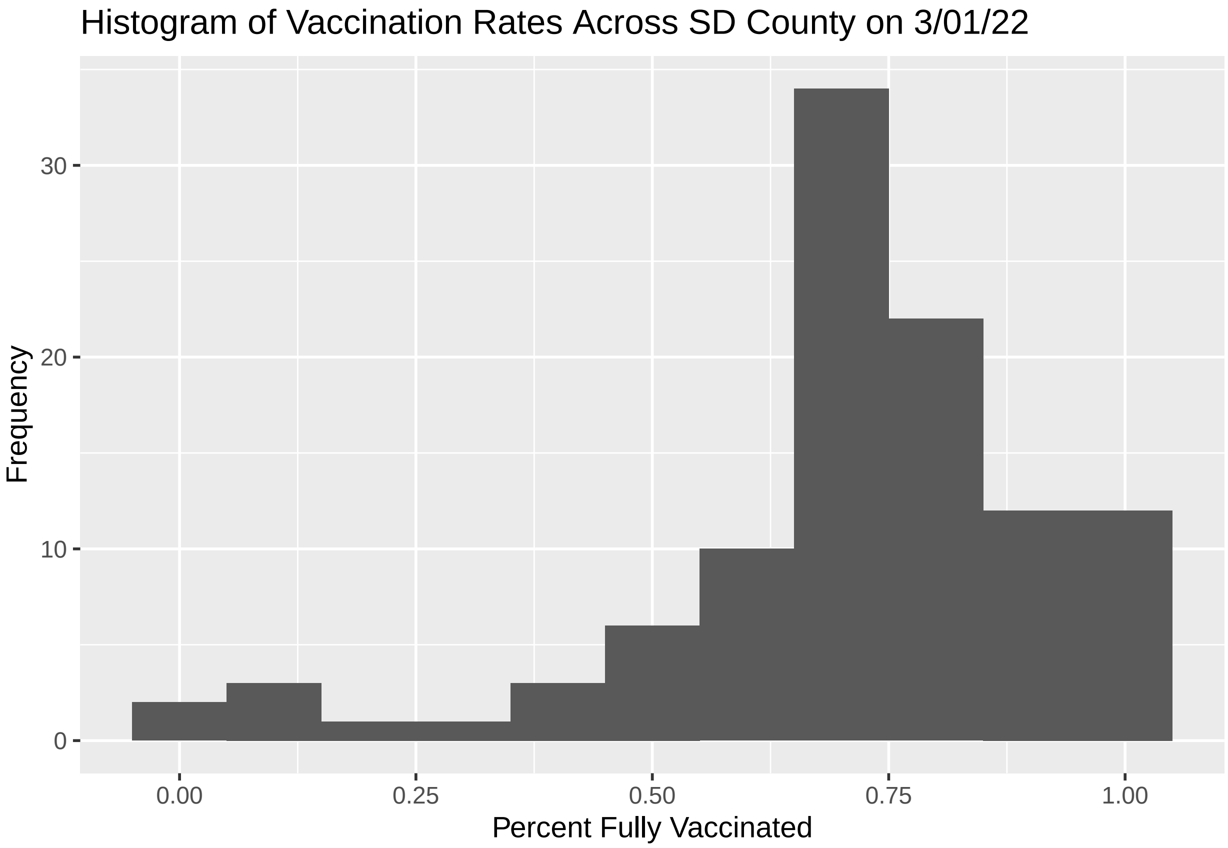
```
sd.n <- sd %>%
  filter(as_of_date == "2022-03-01")
twoday <- na.omit(sd.n$percent_of_population_fully_vaccinated)
mean(sd.n$percent_of_population_fully_vaccinated, na.rm = T)
```

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

On 3/01/2022, the mean vaccination rate in SD county was 70.5%. Plot the vaccination percentages in a histogram.

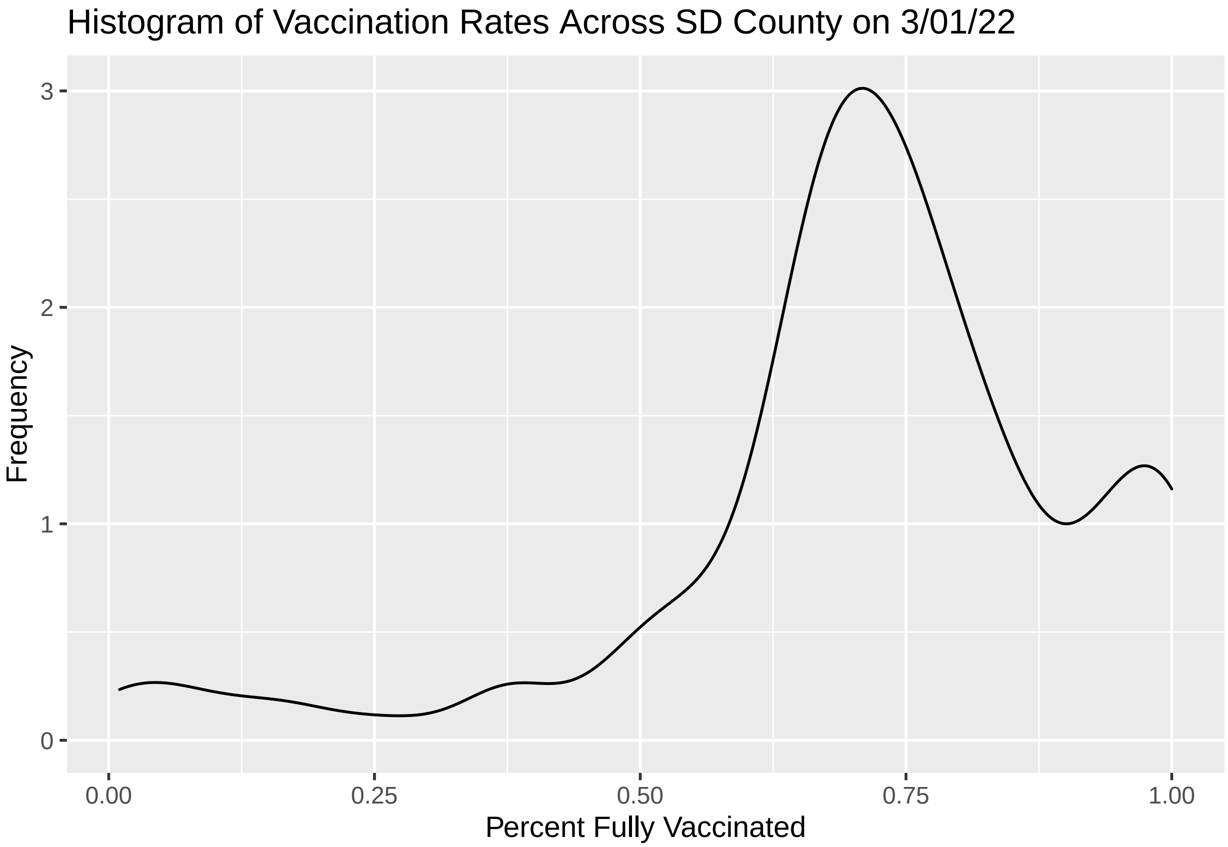
```
ggplot(sd.n, aes(x = percent_of_population_fully_vaccinated)) +
  geom_histogram(binwidth = 0.1) +
  xlab("Percent Fully Vaccinated")+
  ylab("Frequency")+
  labs( title = "Histogram of Vaccination Rates Across SD County on 3/01/22")
```

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



```
ggplot(sd.n, aes(x = percent_of_population_fully_vaccinated)) +
  geom_density() +
  xlab("Percent Fully Vaccinated")+
  ylab("Frequency")+
  labs( title = "Histogram of Vaccination Rates Across SD County on 3/01/22")
```

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

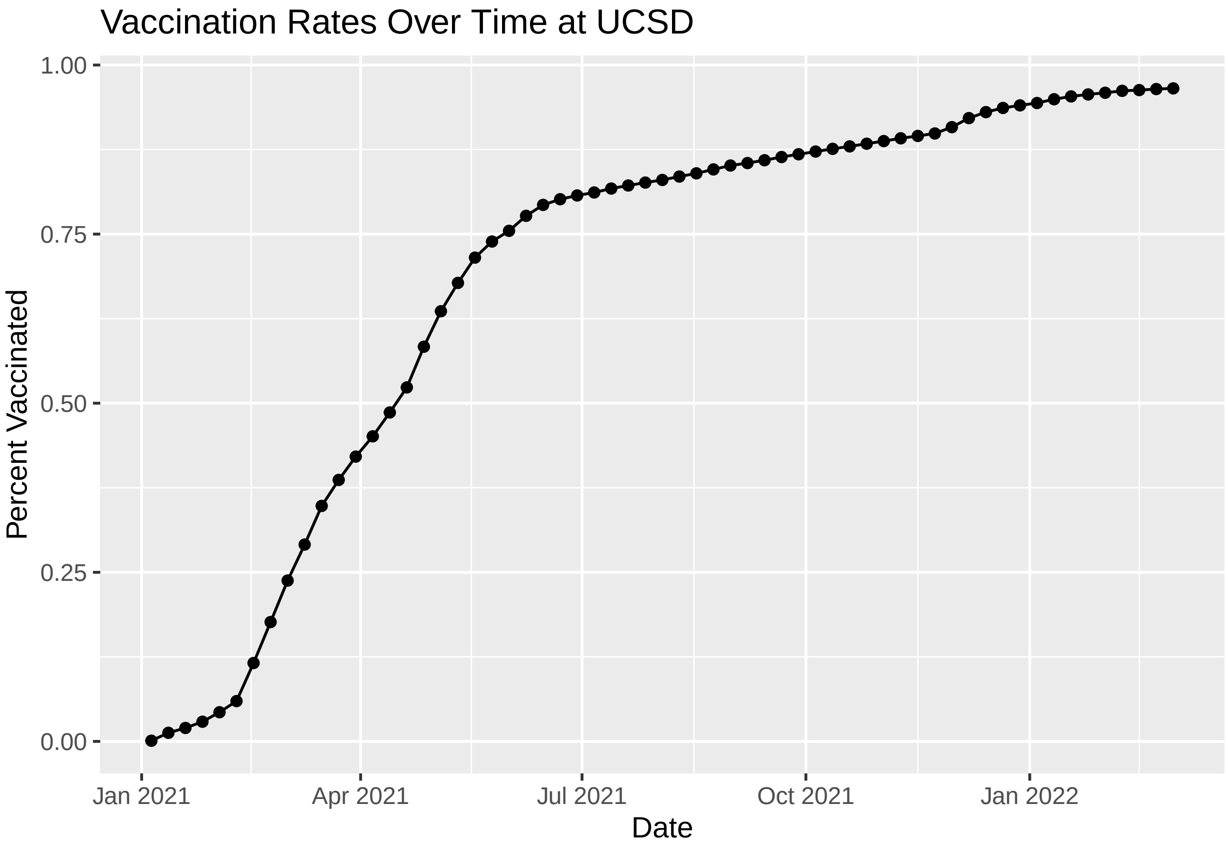


Now focus on UCSD.

```
ucsd <- sd %>% filter(zip_code_tabulation_area == 92037)
```

Vaccination Rate over time for UCSD

```
ggplot(ucsd)+
  aes(x = as_of_date, y = percent_of_population_fully_vaccinated) +
  geom_point()+
  geom_line() +
  labs(title = "Vaccination Rates Over Time at UCSD",
       x = "Date", y = "Percent Vaccinated")
```

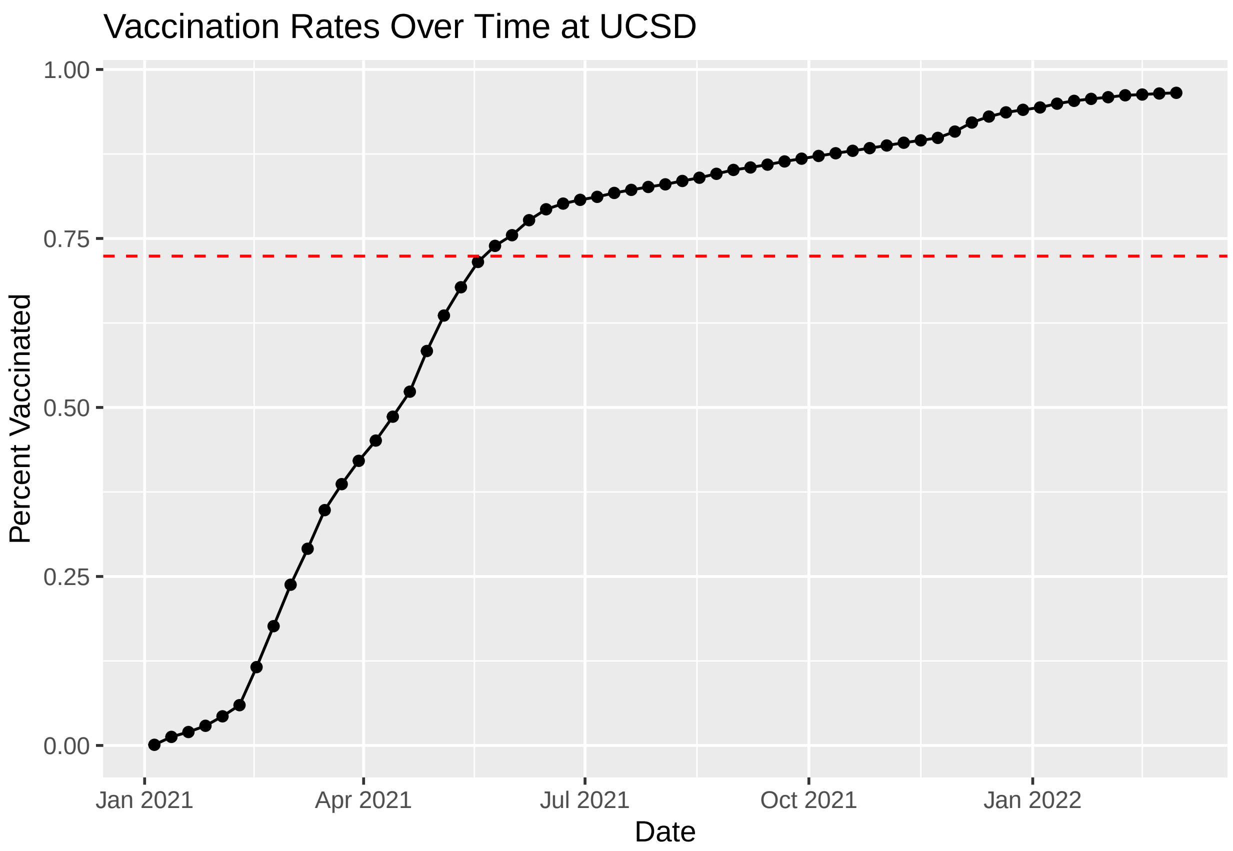


Compare Areas of similar population to UCSD

```
ucsd_size <- unique(ucsd$age5_plus_population)
sd_sim <- sd %>%
  filter(age5_plus_population > ucsd_size & as_of_date == "2022-03-01")
avg_vax <- mean(sd_sim$percent_of_population_fully_vaccinated, na.rm = T)
```

Add a line into the plot from above that is where the average vaccination rate for all ZIP codes at least as large as UCSD.

```
ggplot(ucsd)+
  aes(x = as_of_date, y = percent_of_population_fully_vaccinated) +
  geom_point()+
  geom_line() +
  geom_hline(yintercept =  avg_vax, linetype = "dashed", color = "red")+
  labs(title = "Vaccination Rates Over Time at UCSD",
       x = "Date", y = "Percent Vaccinated")
```



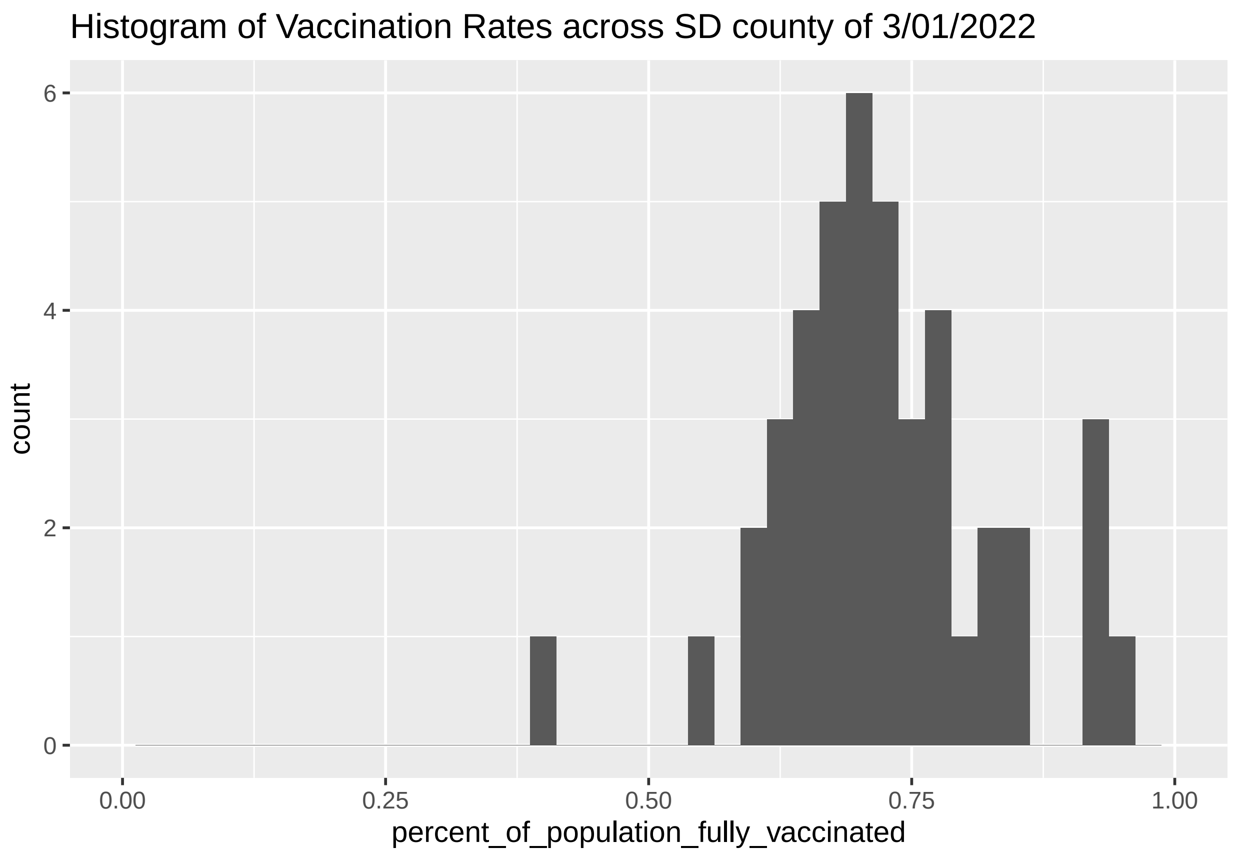
```
summary(sd_sim$percent_of_population_fully_vaccinated)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.3890  0.6618  0.7112  0.7239  0.7793  0.9480
```

The average percent vaccinated is 72.39% for ZIP codes similar to UCSD

```
ggplot(sd_sim, aes(x = percent_of_population_fully_vaccinated)) +
  geom_histogram(binwidth = 0.025)+
  xlim(0,1) +
  labs(title = "Histogram of Vaccination Rates across SD county of 3/01/2022",
       s = "Percent Vaccinated")
```

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



```r
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
```

92040 is below the average of Percent vaccinated in ZIP Codes similarily sized to UCSD.

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

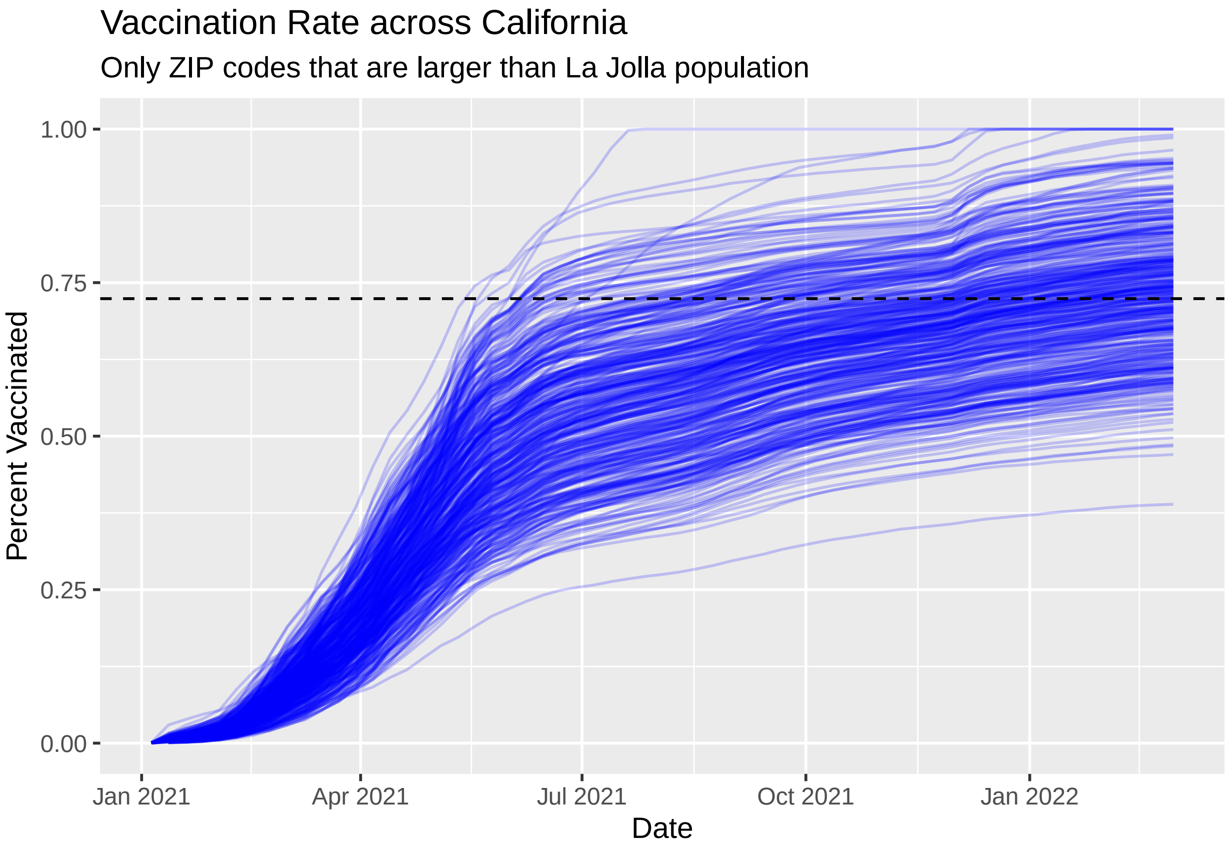
```
##   percent_of_population_fully_vaccinated
## 1                               0.723778
```

92109 is also below the average, but only slightly.

```r
vax_sim_all <- vax %>% filter(age5_plus_population > 36144)


ggplot(vax_sim_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 ZIP codes that are larger than La Jolla population") +
  geom_hline(yintercept = avg_vax, linetype= "dashed")
```

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


Vaccination Rate across California
Only ZIP codes that are larger than La Jolla population
Percent Vaccinated
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
Date