lab17

Zijing

2022-11-27

```
library(skimr)
```

```
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?

2021-01-05

Q4. What is the latest date in this dataset?

2022-11-22

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

Table 1: Data summary

Name	vax

Table 1: Data summary

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	13685.10	31756.12	88556.7	
age5_plus_population	0	1.00	20875.24	21105.98	0	1460.50	15364.00	34877.00	101902.0	
tot_population	8514	0.95	23372.77	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	87186.0	
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	
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	12309.19	14555.83	0	466.00	5810.00	21140.00	86696.0	

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?

14921

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

8.54

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

Missing data from some zip code areas.

```
library(lubridate)
```

```
## Loading required package: timechange

##
## Attaching package: 'lubridate'

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

```
today()
```

```
## [1] "2022-11-26"
```

```
vax$as_of_date <- ymd(vax$as_of_date)
```

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

```
## Time difference of 690 days
```

```
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?

4 days

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

99 unique dates

```
library(zipcodeR)
```

```
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, ...
```

```
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
```

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?

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

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

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

```
sd_221115 <- filter(sd, as_of_date == "2022-11-15")
mean(sd_221115$percent_of_population_fully_vaccinated, na.rm = TRUE)
```

```
## [1] 0.7369099
```

Average percentage is 73.7%

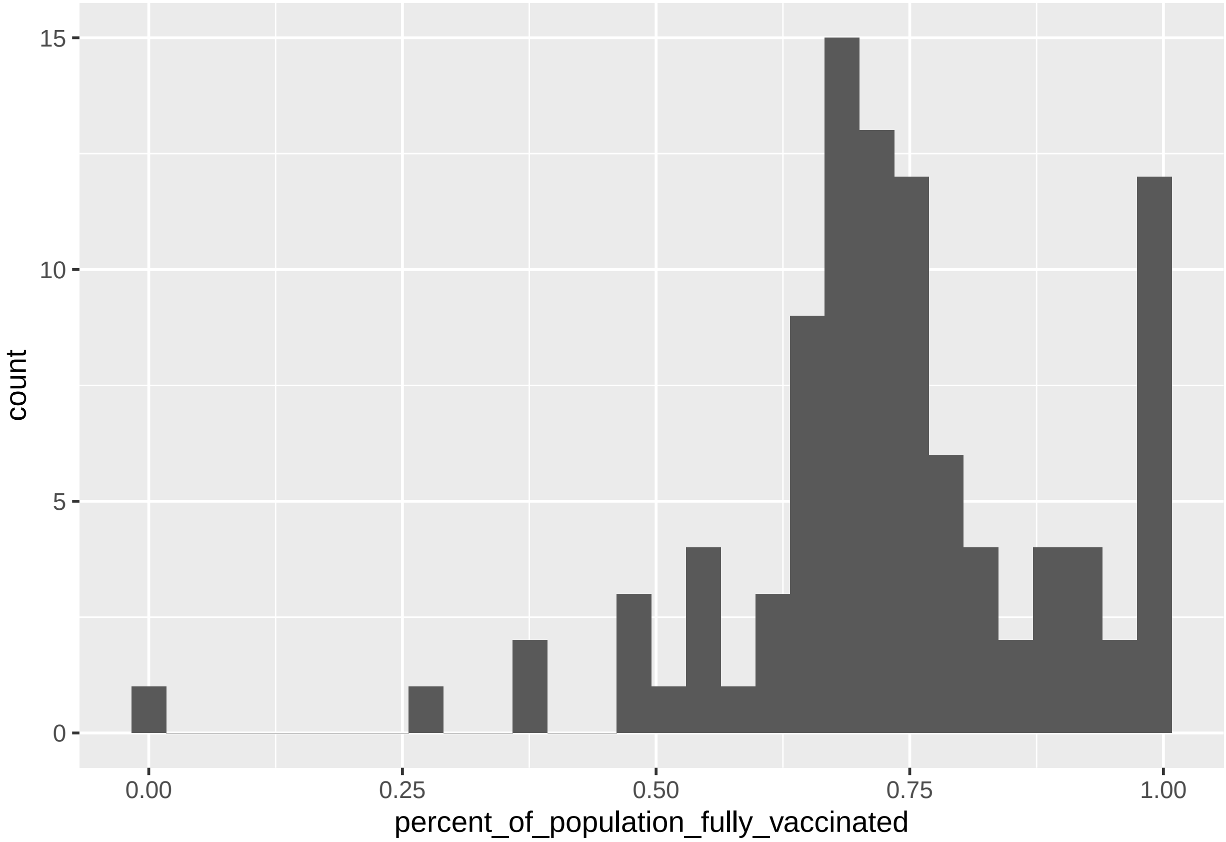
Q14. Using either ggplot or base R graphics make a summary figure that shows the distribution of Percent of Population Fully Vaccinated values as of "2022-11-15"?

```
library(ggplot2)
```

```
ggplot(sd_221115, aes(percent_of_population_fully_vaccinated))+
  geom_histogram()
```

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

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

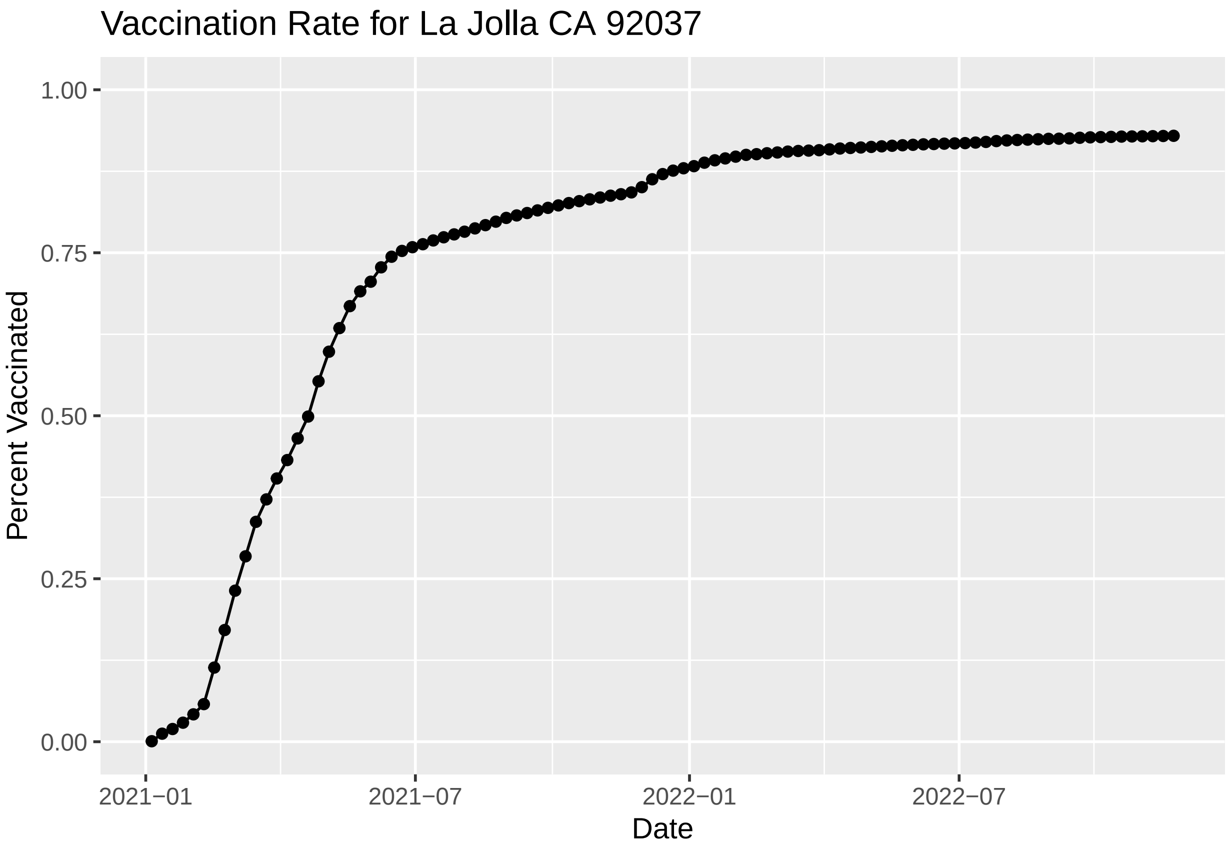


```
ucsd <- filter(sd, zip_code_tabulation_area=="92037")
ucsd[1,]$age5_plus_population
```

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

Q15

```
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",title="Vaccination Rate for La Jolla CA 92037")
```



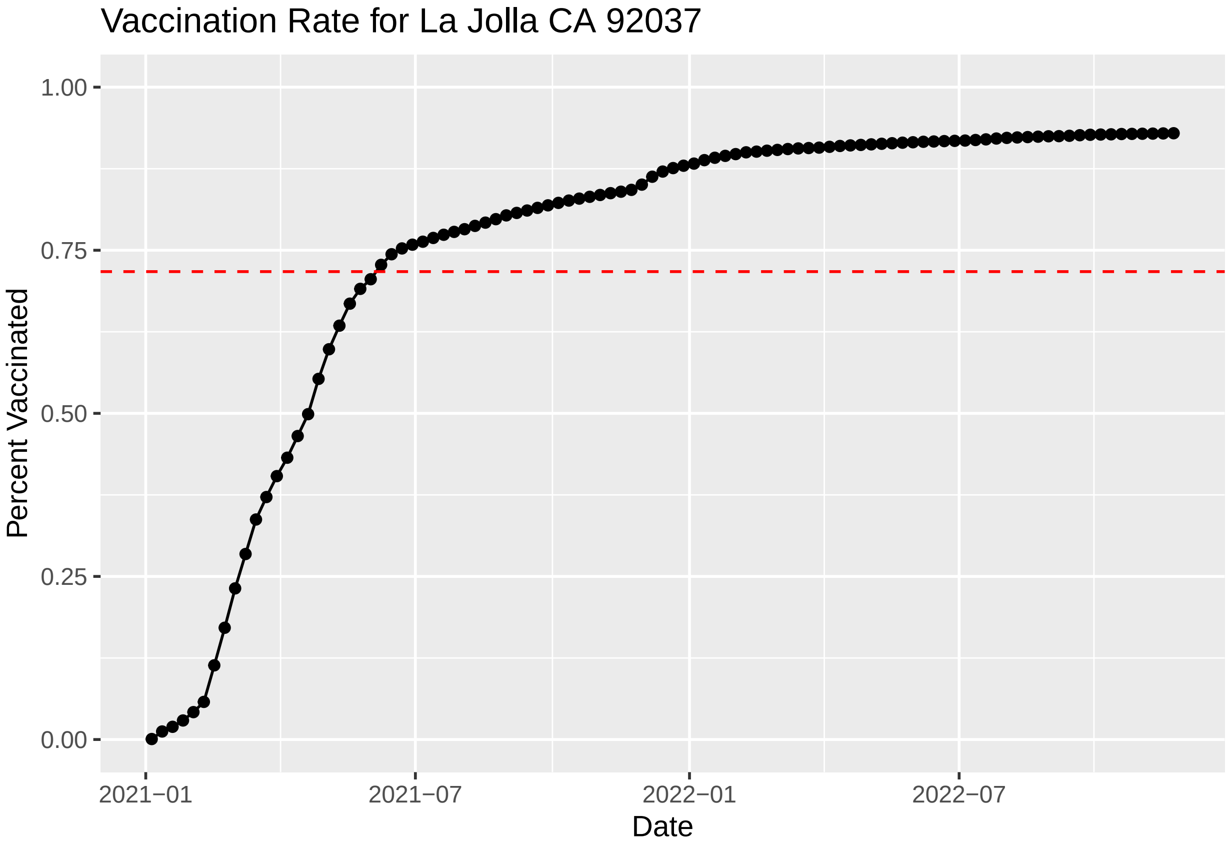
```
vax.36 <- filter(vax, age5_plus_population > 36144 &
                 as_of_date == "2022-11-15")
```

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

```
## [1] 0.7172851
```

Q16

```
ggplot(ucsd) +
  aes(as_of_date,
      percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) +
  geom_hline(yintercept = comp_mean, linetype = "dashed", color="red")+
  ylim(c(0,1)) +
  labs(x="Date", y="Percent Vaccinated",title="Vaccination Rate for La Jolla CA 92037")
```



# 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"?

```
skimr::skim(vax.36)
```

Table 4: Data summary

Name	vax.36
Number of rows	411
Number of columns	18
Column type frequency:	
character	4
Date	1
numeric	13

Table 4: Data summary

Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
local_health_jurisdiction	0	1	4	15	0	37	0
county	0	1	4	15	0	36	0
vem_source	0	1	26	26	0	1	0
redacted	0	1	2	2	0	1	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
as_of_date	0	1	2022-11-15	2022-11-15	2022-11-15	1

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
zip_code_tabulation_area	0	1	92862.10	1716.60	90001.00	91761.50	92646.00	94517.00	96003.00	
vaccine_equity_metric_quartile	0	1	2.35	1.11	1.00	1.00	2.00	3.00	4.00	
age12_plus_population	0	1	46847.40	12057.32	31650.90	37693.55	43985.40	53931.50	88556.70	
age5_plus_population	0	1	52012.32	13620.19	36181.00	41612.50	48573.00	59167.50	101902.00	
tot_population	0	1	55640.91	14745.19	38007.00	44393.00	52212.00	62910.00	111165.00	
persons_fully_vaccinated	0	1	39837.28	11739.80	17422.00	31926.50	37064.00	45033.50	87151.00	
persons_partially_vaccinated	0	1	4077.70	2620.74	1733.00	2813.00	3542.00	4666.00	39160.00	
percent_of_population_fully_vaccinated	0	1	0.72	0.11	0.38	0.64	0.72	0.79	1.00	
percent_of_population_partially_vaccinated	0	1	0.07	0.05	0.04	0.06	0.06	0.08	0.98	
percent_of_population_with_1_plus_dose	0	1	0.79	0.11	0.44	0.71	0.79	0.86	1.00	
booster_recip_count	0	1	22817.37	7812.12	8603.00	17134.50	21640.00	27265.50	56744.00	
bivalent_dose_recip_count	0	1	5618.65	2952.70	1375.00	3418.50	4941.00	7269.50	16829.00	
eligible_recipient_count	0	1	39609.31	11653.38	17321.00	31819.50	36758.00	44903.50	86696.00	

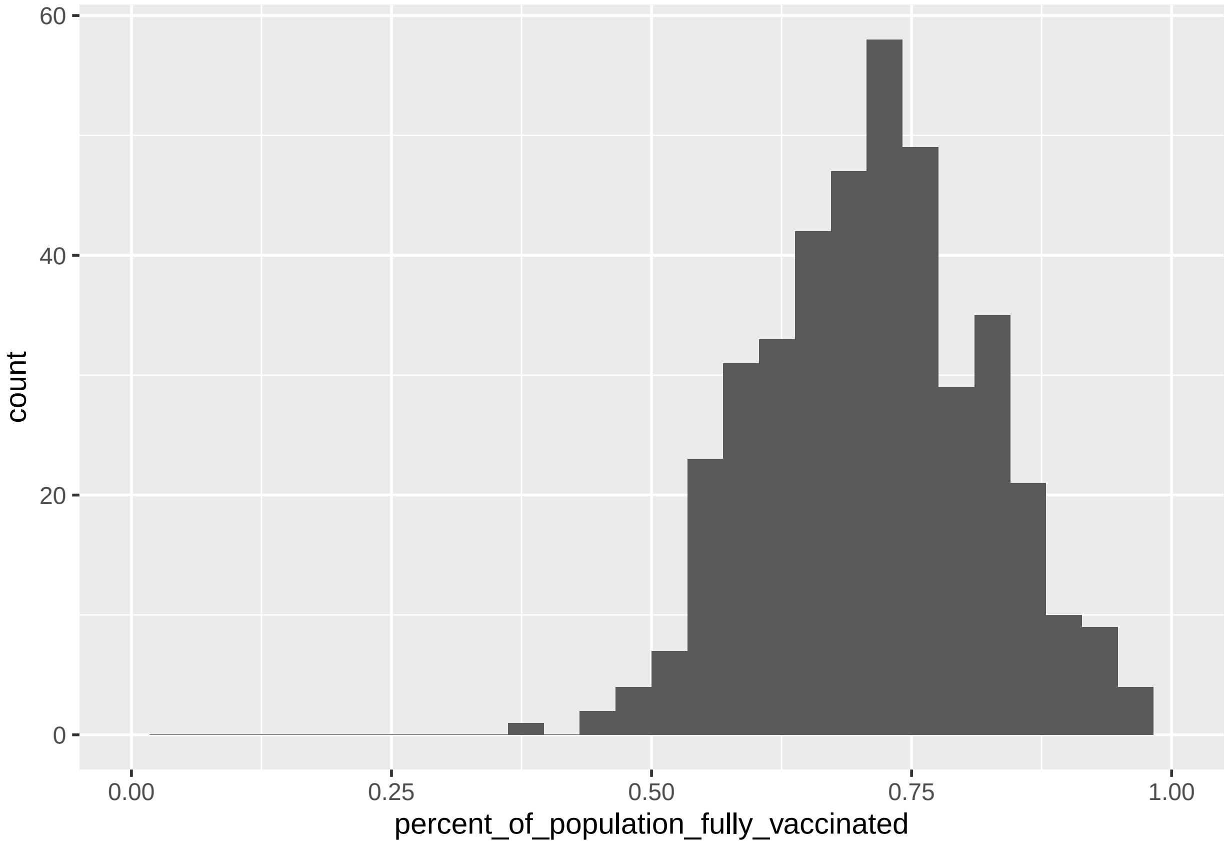
Min 3.78501e-01, 1st Qu. 6.396185e-01, Median 7.15524e-01, Mean 7.172851e-01, 3rd Qu. 7.879820e-01, and Max 1.00000e+00

Q18

```
ggplot(vax.36, aes(percent_of_population_fully_vaccinated))+
  geom_histogram()+
  xlim(0,1)
```

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

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



# 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
```

```
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
```

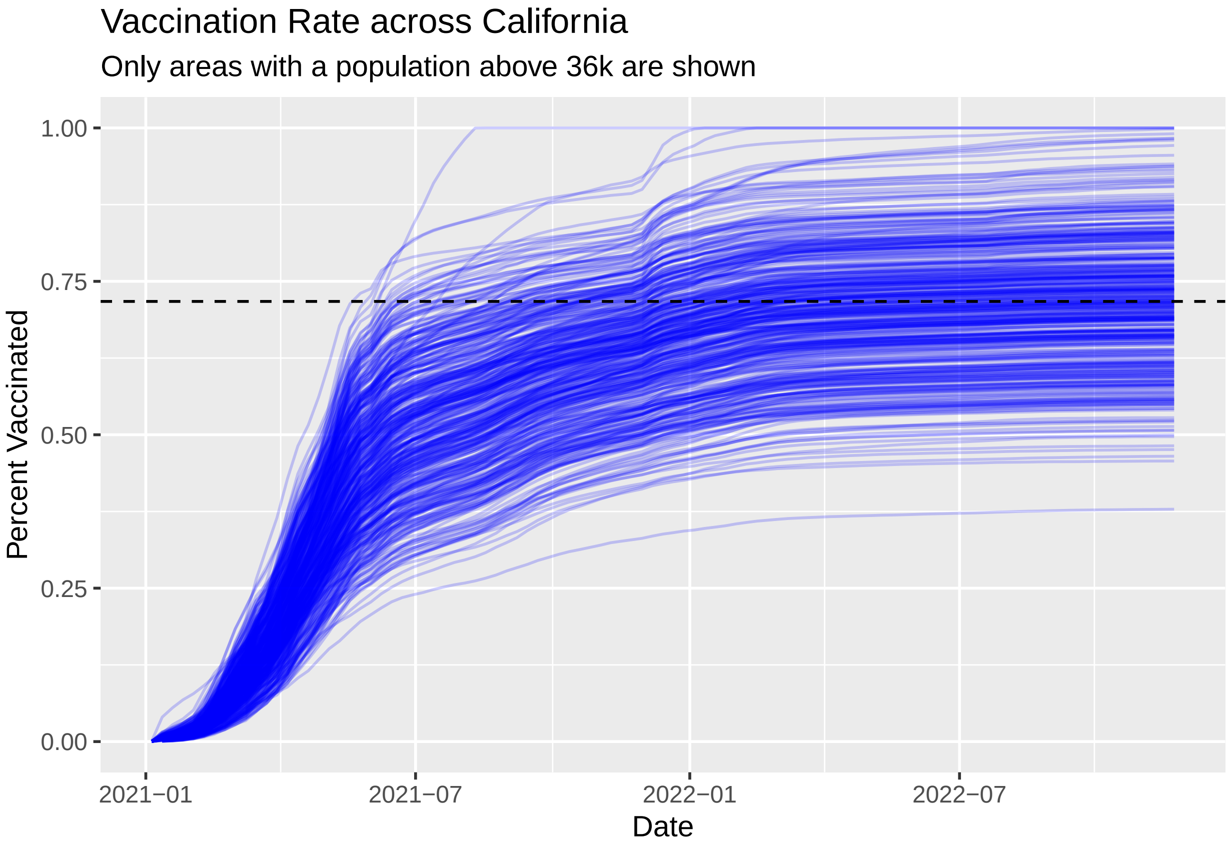
Both areas are below the average calculated above.

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)


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(0.00,1.00) +
  labs(x="Date", y="Percent Vaccinated",
       title="Vaccination Rate across California",
       subtitle="Only areas with a population above 36k are shown") +
  geom_hline(yintercept = comp_mean, linetype="dashed")
```

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


Vaccination Rate across California
Only areas with a population above 36k are shown
Percent Vaccinated
Date