Class 17: Vaccination Rate Mini-Project

Michelle Woo

Importing the data

```
# importing the 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                   93704                   Fresno     Fresno
2 2021-01-05                   95684                El Dorado  El Dorado
3 2021-01-05                   92273                 Imperial   Imperial
4 2021-01-05                   93662                   Fresno     Fresno
5 2021-01-05                   95673               Sacramento Sacramento
6 2021-01-05                   93668                   Fresno     Fresno
  vaccine_equity_metric_quartile                vem_source
1                              1 Healthy Places Index Score
2                              2 Healthy Places Index Score
3                              1 Healthy Places Index Score
4                              1 Healthy Places Index Score
5                              2 Healthy Places Index Score
6                              1    CDPH-Derived ZCTA Score
  age12_plus_population age5_plus_population tot_population
1               24803.5                27701          29740
2                2882.9                 3104           3129
3                1633.1                 1763           2010
4               24501.3                28311          30725
5               13671.7                15453          16636
6                1013.4                 1199           1219
  persons_fully_vaccinated persons_partially_vaccinated
1                       NA                           NA
2                       NA                           NA
```

```
3                       NA                            NA
4                       NA                            NA
5                       NA                            NA
6                       NA                            NA
  percent_of_population_fully_vaccinated
1                                     NA
2                                     NA
3                                     NA
4                                     NA
5                                     NA
6                                     NA
  percent_of_population_partially_vaccinated
1                                         NA
2                                         NA
3                                         NA
4                                         NA
5                                         NA
6                                         NA
  percent_of_population_with_1_plus_dose booster_recip_count
1                                     NA                  NA
2                                     NA                  NA
3                                     NA                  NA
4                                     NA                  NA
5                                     NA                  NA
6                                     NA                  NA
  bivalent_dose_recip_count eligible_recipient_count
1                        NA                        5
2                        NA                        0
3                        NA                        1
4                        NA                        1
5                        NA                        3
6                        NA                        0
  eligible_bivalent_recipient_count
1                                 5
2                                 0
3                                 0
4                                 1
5                                 3
6                                 0
                                                                                                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?

```
colnames(vax)
```

```
 [1] "as_of_date"
 [2] "zip_code_tabulation_area"
 [3] "local_health_jurisdiction"
 [4] "county"
 [5] "vaccine_equity_metric_quartile"
 [6] "vem_source"
 [7] "age12_plus_population"
 [8] "age5_plus_population"
 [9] "tot_population"
[10] "persons_fully_vaccinated"
[11] "persons_partially_vaccinated"
[12] "percent_of_population_fully_vaccinated"
[13] "percent_of_population_partially_vaccinated"
[14] "percent_of_population_with_1_plus_dose"
[15] "booster_recip_count"
[16] "bivalent_dose_recip_count"
[17] "eligible_recipient_count"
[18] "eligible_bivalent_recipient_count"
[19] "redacted"
```

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?

2023-05-23

Using skim() to get an overview

```
skimr::skim_without_charts(vax)
```

Table 1: Data summary

Name	vax
Number of rows	222264
Number of columns	19
_____	
Column type frequency:	
character	5
numeric	14
_____	
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	126	0
local_health_jurisdiction	0	1	0	15	630	62	0
county	0	1	0	15	630	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
zip_code_tabulation_area	0	1.00	93665.11	817.38	90001	92257.75	93658.50	95380.50	97635.0
vaccine_equity_metric_quartile	10962	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	20875.22	21105.96	0	1460.50	15364.00	34877.00	101902.0
tot_population	10836	0.95	23372.72	22628.50	12	2126.00	18714.00	38168.00	111165.0
persons_fully_vaccinated	17848	0.92	14299.49	15281.94	11	957.00	9034.00	23818.00	87721.0
persons_partially_vaccinated	17848	0.92	1712.08	2075.03	11	164.00	1204.00	2551.00	43152.0
percent_of_population_fully_vaccinated	22720	0.90	0.58	0.25	0	0.44	0.62	0.75	1.0
percent_of_population_partially_vaccinated	22720	0.90	0.08	0.09	0	0.05	0.06	0.08	1.0
percent_of_population_with_1_plus_dose	22883	0.90	0.65	0.24	0	0.50	0.68	0.82	1.0
booster_recip_count	74543	0.66	6417.22	7795.13	11	331.00	3135.00	10344.00	60058.0

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
bivalent_dose_recip_count	160089	0.28	3438.22	4034.61	11	225.00	1863.00	5532.00	29593.0
eligible_recipient_count	0	1.00	13145.14	15144.22	0	537.00	6691.00	22558.00	87442.0
eligible_bivalent_recipient_count	0	1.00	13038.24	15218.39	0	263.00	6583.00	22550.00	87442.0

Q5. How many numeric columns are in the dataset?

```
numeric_columns <- sapply(vax, is.numeric)
num_numeric_columns <- sum(numeric_columns)
num_numeric_columns
```

```
[1] 14
```

Q6. How many `NA` values are there in the `persons_fully_vaccinated` column?

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

```
[1] 17848
```

17711 (previous dataset)

17848 (updated)

Q7. What percent of `persons_fully_vaccinated` values are missing?

```
# finding the total value
total_value <- nrow(vax)

# taken from Q6.
num_na_values <- sum(is.na(vax$persons_fully_vaccinated))

# finding the percentage
percentage_missing <- (num_na_values / total_value) * 100

percentage_missing
```

```
[1] 8.03009
```

8.03%

Q8. Why might this data be missing?

The data might be missing because it is difficult to collect all data from many people for reasons such as having confidential restrictions, inaccurate census data, or difficulty in reaching all sorts of people in all of the state, not just the major cities.

Working with dates

```
# using lubridate to format our data
library(lubridate)
```

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

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

    date, intersect, setdiff, union
```

```
today()
```

```
[1] "2023-05-31"
```

```
# year-month-day format
vax$as_of_date <- ymd(vax$as_of_date)
```

With the new format, we can do math with the dates. Such as answering the question: how many days have passed since the first vaccination reported in the dataset?:

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

```
Time difference of 876 days
```

And days the dataset spans:

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

```
Time difference of 875 days
```

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

```
# from the last row of the date to today
today() - ymd(vax$as_of_date[nrow(vax)])
```

```
Time difference of 1 days
```

8 day (from previous dataset)

1 day (updated)

Q10: How many unique dates are in the dataset (how many different dates are detailed)?

```
# using length to count the vector
length(unique(vax$as_of_date))
```

```
[1] 126
```

125 unique dates (in old dataset)

126 (updated)

Working with ZIP codes

Using zipcodeR:

```
# install.packages('zipcodeR')

#loading it in
library(zipcodeR)
```

```
geocode_zip('92037')
```

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

```
# distance between two zipcodes
zip_distance('92037','92109')
```

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

```
# pulling census data
reverse_zipcode(c('92037','92109'))
```

```
# A tibble: 2 x 24
  zipcode zipcode_type major_city post_office_city common_city_list county state
  <chr>   <chr>        <chr>      <chr>                     <blob> <chr>  <chr>
1 92037   Standard     La Jolla   La Jolla, CA           <raw 20 B> San D~ CA
2 92109   Standard     San Diego  San Diego, CA          <raw 21 B> San D~ CA
# i 17 more variables: lat <dbl>, lng <dbl>, timezone <chr>,
#   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>
```

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

Focusing on the SD area

```
# subset to SD area
sd <- vax[vax$county == "San Diego", ]
```

or using dplyr:

```
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] 13482
```

Q11. How many distinct zip codes are listed in the SD County?

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

```
[1] 107
```

Q12. What SD County Zip code area has the largest population in this dataset?

```
largest <- sd[sd$age5_plus_population == max(sd$age5_plus_population),]

unique(largest$zip_code_tabulation_area)
```

```
[1] 92154
```

ZIP code: 92154

Q13. What is the overall average for all SD county as of 2023-05-23?

```
# using tidyverse approach
sd_may23 <- filter(sd, as_of_date == '2023-05-23')

# 107 zip codes
dim(sd_may23)
```

```
[1] 107  19
```

```
# finding the mean
mean(sd_may23$percent_of_population_fully_vaccinated, na.rm = T)
```

```
[1] 0.7419992
```

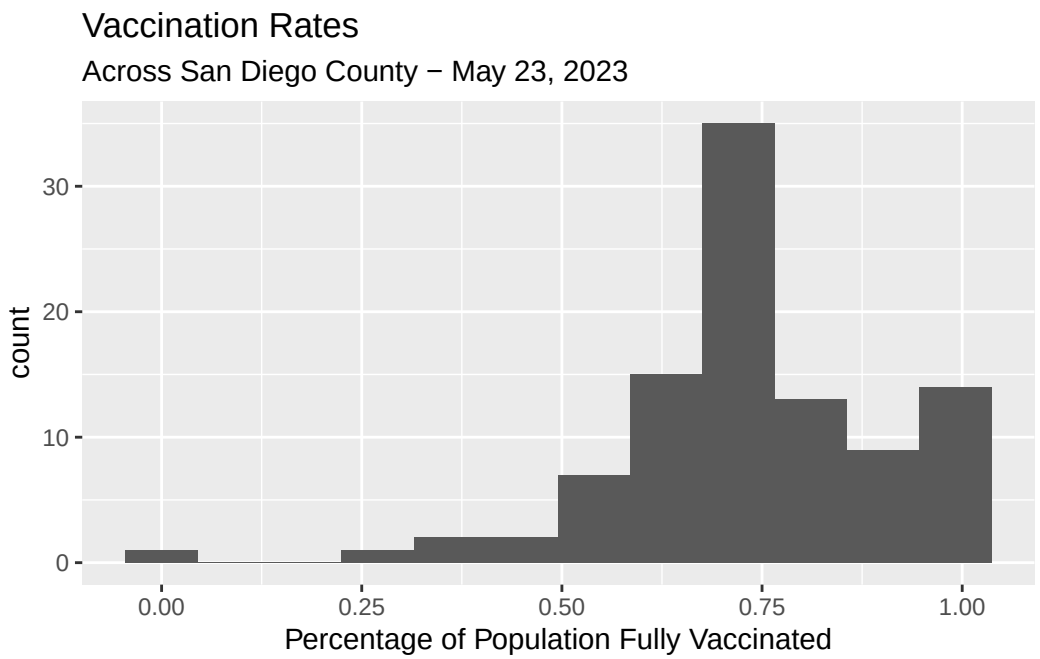
An average of 0.74 / 74% of people are fully vaccinated in the SD county.

Q14. Make a summary figure that shows the distribution of Percent of Population Fully Vaccinated values as of 2023-05-23

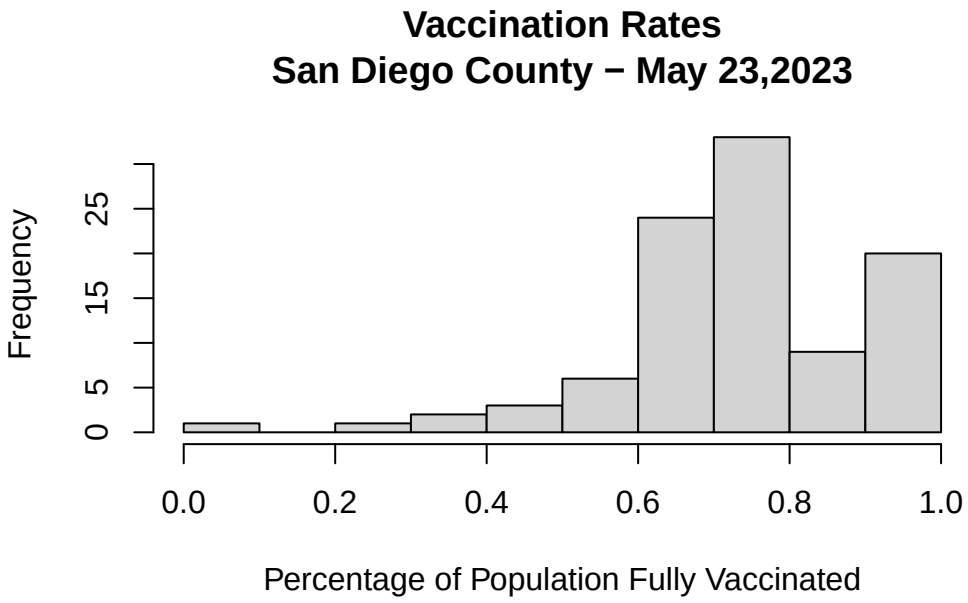
```
# ggplot
library(ggplot2)

ggplot(sd_may23) +
  aes(percent_of_population_fully_vaccinated) +
  geom_histogram(bins = 12) +
  labs(title = 'Vaccination Rates',
       subtitle = 'Across San Diego County - May 23, 2023',
       x = 'Percentage of Population Fully Vaccinated')
```

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



```r
# base r
hist(sd_may23$percent_of_population_fully_vaccinated,
     xlab = 'Percentage of Population Fully Vaccinated',
     main = 'Vaccination Rates\nSan Diego County - May 23,2023')
```



Focusing on UCSD/La Jolla

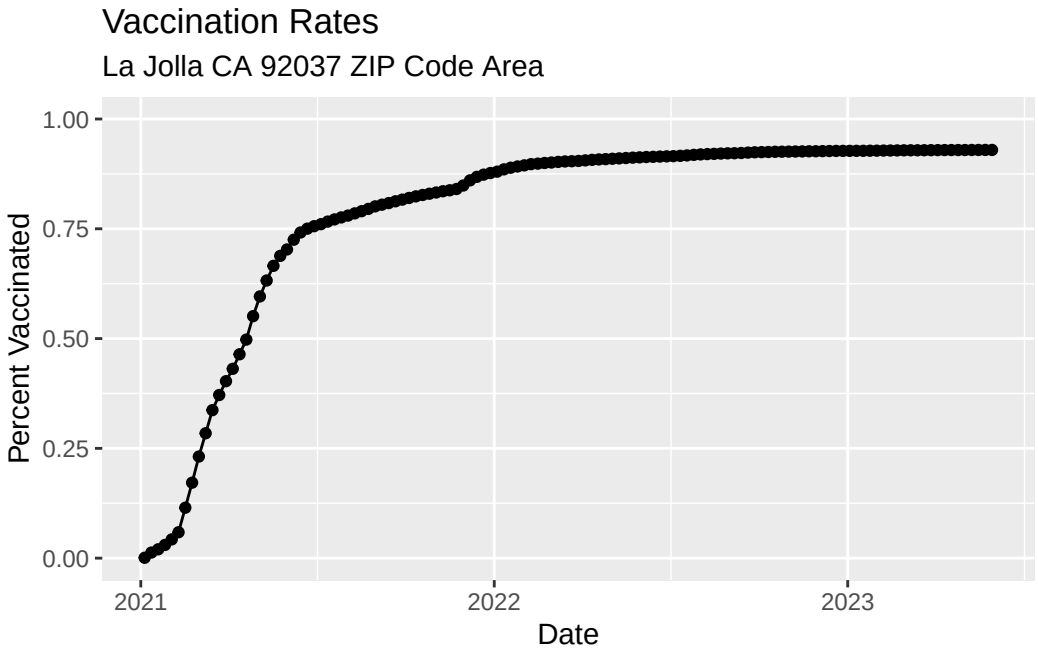
```r
ucsd <- filter(sd, zip_code_tabulation_area=="92037")
ucsd[1,]$age5_plus_population
```

```
[1] 36144
```

Q15. Using ggplot, make a graph of the vaccination rate time course for the 92037 ZIP code area

```r
ggplot(ucsd) +
  aes(as_of_date,
      percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group = 1) +
```

```
  ylim(c(0,1)) +
  labs(title = 'Vaccination Rates',
       subtitle = 'La Jolla CA 92037 ZIP Code Area',
       x = 'Date',
       y = 'Percent Vaccinated')
```



Comparing to similar sized areas

Let's return to the full dataset and look across every zip code area with a population at least as large as that of 92037 on *as_of_date* "2023-05-23"

```
# Subset to all CA areas with a population as large as 92037
vax.36 <- filter(vax, age5_plus_population > 36144 &
                 as_of_date == "2023-05-23")

head(vax.36)
```

```
  as_of_date zip_code_tabulation_area local_health_jurisdiction      county
1 2023-05-23                   90805                Long Beach Los Angeles
2 2023-05-23                   93257                    Tulare      Tulare
3 2023-05-23                   90004               Los Angeles Los Angeles
```

```
4 2023-05-23                    90808               Long Beach Los Angeles
5 2023-05-23                    95355               Stanislaus  Stanislaus
6 2023-05-23                    90802               Long Beach Los Angeles
  vaccine_equity_metric_quartile                vem_source
1                              1 Healthy Places Index Score
2                              1 Healthy Places Index Score
3                              1 Healthy Places Index Score
4                              4 Healthy Places Index Score
5                              2 Healthy Places Index Score
6                              1 Healthy Places Index Score
  age12_plus_population age5_plus_population tot_population
1               77165.9                88279          95995
2               61519.8                70784          76519
3               52412.5                57024          60541
4               33952.3                37179          39330
5               50941.6                56248          59621
6               35238.1                37017          38962
  persons_fully_vaccinated persons_partially_vaccinated
1                    62829                         6949
2                    45117                         5629
3                    47272                         5963
4                    30283                         2375
5                    39616                         3210
6                    28152                         3711
  percent_of_population_fully_vaccinated
1                               0.654503
2                               0.589618
3                               0.780826
4                               0.769972
5                               0.664464
6                               0.722550
  percent_of_population_partially_vaccinated
1                                   0.072389
2                                   0.073563
3                                   0.098495
4                                   0.060386
5                                   0.053840
6                                   0.095247
  percent_of_population_with_1_plus_dose booster_recip_count
1                               0.726892               33175
2                               0.663181               22223
3                               0.879321               29130
4                               0.830358               20463
```

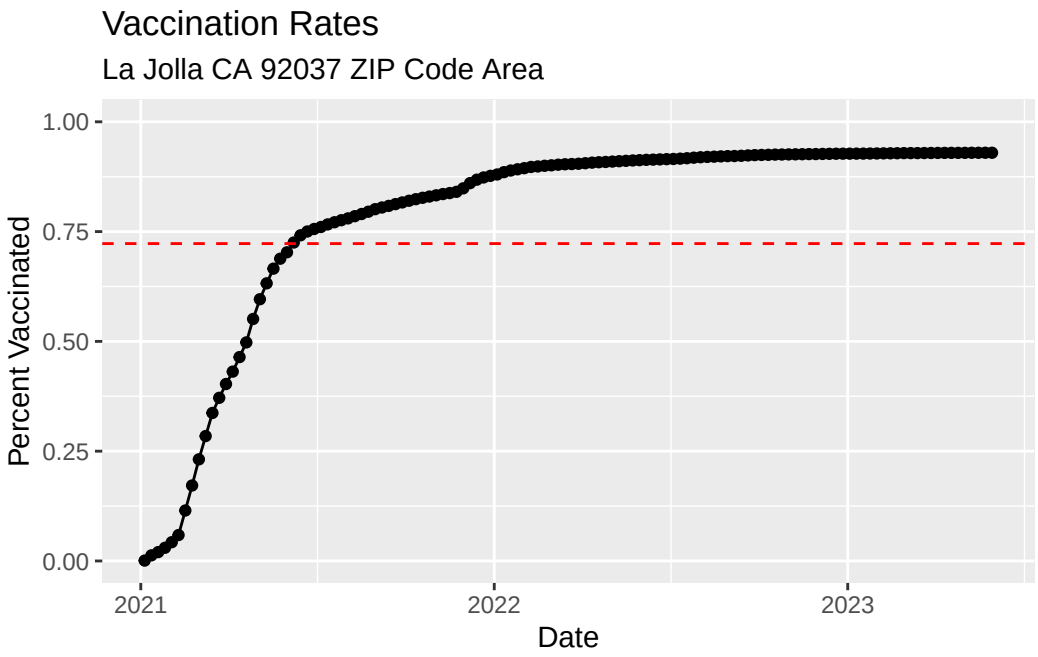
```
5                              0.718304                22873
6                              0.817797                17033
  bivalent_dose_recip_count eligible_recipient_count
1                     10919                    62713
2                      5297                    45104
3                     12081                    47148
4                      9676                    30203
5                      8291                    39588
6                      7169                    28107
  eligible_bivalent_recipient_count redacted
1                             62713       No
2                             45104       No
3                             47148       No
4                             30203       No
5                             39588       No
6                             28107       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* "2023-05-23". Add this as a straight horizontal line to your plot from above with the `geom_hline()` function:

```r
# ucsd ggplot in a variable
plot <- ggplot(ucsd) +
  aes(x = as_of_date,
      y = percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group = 1) +
  ylim(c(0,1)) +
  labs(title = 'Vaccination Rates',
       subtitle = 'La Jolla CA 92037 ZIP Code Area',
       x = 'Date',
       y = 'Percent Vaccinated')

# mean of 0.723
mean_vax.36 <- mean(vax.36$percent_of_population_fully_vaccinated)

# adding on the mean to previous ggplot
plot + geom_hline(yintercept = mean_vax.36, linetype = 'dashed', color = 'red')
```



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* "2023-05-23"?

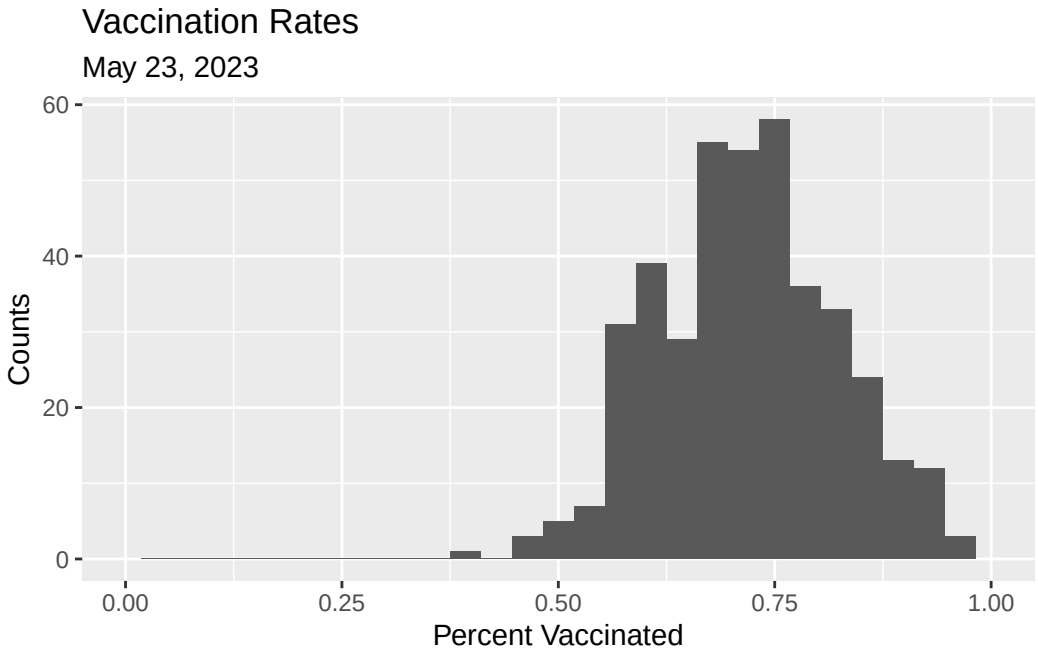
```
summary(vax.36$percent_of_population_fully_vaccinated)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.3815  0.6470  0.7208  0.7227  0.7923  1.0000 
```

Q18. Using ggplot, generate a histogram of this data:

```
ggplot(vax.36) +
   aes(percent_of_population_fully_vaccinated) +
  geom_histogram(bins = 29) +
  xlim(c(0,1)) +
  labs(title = 'Vaccination Rates',
       subtitle = 'May 23, 2023',
       x = 'Percent Vaccinated',
       y= 'Counts')
```

```
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?

```
# 92040 code
vax %>% filter(as_of_date == "2023-05-23") %>%
  filter(zip_code_tabulation_area=="92040") %>%
  select(percent_of_population_fully_vaccinated)
```

```
  percent_of_population_fully_vaccinated
1                               0.552499
```

```
# 92109 code
vax %>% filter(as_of_date == "2023-05-23") %>%
  filter(zip_code_tabulation_area=="92109") %>%
  select(percent_of_population_fully_vaccinated)
```

```
  percent_of_population_fully_vaccinated
1                               0.694763
```

Both below

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='darkolivegreen4') +
  ylim(0,1) +
  labs(x = 'Date', y = 'Percent Vaccinated',
       title = 'Vaccination Rate Across California',
       subtitle = 'Only areas with population above 36k are shown') +
  geom_hline(yintercept = 0.722, linetype = 'dashed', color = 'forestgreen')
```

```
Warning: Removed 185 rows containing missing values (`geom_line()`).
```