Vaccination Rates Mini-Project

Background

The goal of this hands-on mini-project is to examine and compare the Covid-19 vaccination rates around San Diego. The main dataset for this project comes from "Statewide COVID-19 Vaccines Administered by ZIP Code" CSV file.

Packages Used in this Project

DPLYR: working with and modification of data *SKIMR*: summaries of data sets *LUBRIDATE*: working with dates (i.e. do math) *zipcodeR*: numeric calculations on zipcodes

```
#Lets import the dataset
library(bio3d)
vax <- read.csv("covid19vaccinesbyzipcode_test.csv")
```

Exploratory Data Analysis

```
#Inspect the dataset
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
```

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? -> *2022-03-01*

```
#More Summary Data
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

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
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	

```r
na.omit(vax[vax$persons_fully_vaccinated == 0,])
```

```
##  [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] persons_fully_vaccinated
## [10] persons_partially_vaccinated
## [11] percent_of_population_fully_vaccinated
## [12] percent_of_population_partially_vaccinated
## [13] percent_of_population_with_1_plus_dose
## [14] booster_recip_count
## [15] redacted
## <0 rows> (or 0-length row.names)
```

```r
#percentage of NA values in persons_fully_vaccinated column
sum(is.na(vax$persons_fully_vaccinated))/nrow(vax)
```

```
## [1] 0.1704212
```

```r
18338/107604
```

```
## [1] 0.1704212
```

Q5. How many numeric columns are in this dataset? -> *9 numeric columns*

Q6. Note that there are “missing values” in the dataset. How many NA values there in the persons_fully_vaccinated column? -> *18338 N/A values in the “persons_fully_vaccinated” column*

Q7. What percent of persons_fully_vaccinated values are missing (to 2 significant figures)? *There are 17% of N/A values in the “persons_fully_vaccinated” column*

Q8. [Optional]: Why might this data be missing? *There are no zero values in the data set, so NA might be being used instead of 0. The military areas around San Diego are also not required to report their vaccination rates to ca.gov.*

Working with Dates

```
library(lubridate)
```

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

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

```
today()
```

```
## [1] "2022-03-03"
```

```
#Must specify that we are using the year-month-day format in the table
vax$as_of_date <- ymd(vax$as_of_date)
today() - vax$as_of_date[1] #running for 422 days
```

```
## Time difference of 422 days
```

```
today() - vax$as_of_date[nrow(vax)] #last update 2 days ago
```

```
## Time difference of 2 days
```

```
length(unique(vax$as_of_date)) #61 unique dates in the data set
```

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

Q9. How many days have passed since the last update of the dataset? -> *2 days*

Q10. How many unique dates are in the dataset (i.e. how many different dates are detailed)? -> *61 unique dates in the dataset as of 03/03/22* <- read that however you want ;)

Working with Zipcodes

```
library(zipcodeR)
#get coordinates of the centroid Of a zip code
geocode_zip('92037')
```

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

```
#calculate distances between centroids of zip codes
zip_distance('92037','92109')
```

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

```
#pull up census data on a zip code
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
## # ... with 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>
```

We can then use packages like *leaflet* and *ggplot* to superimpose this data onto maps to produce a useful graphical summary.

Focus in on San Diego

```
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")
length(unique(sd$zip_code_tabulation_area)) #107 unique zip codes in San Diego county
```

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

```
sd[which.max(sd$age12_plus_population),]
```

```
##    as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 91 2021-01-05                    92154                 San Diego San Diego
##    vaccine_equity_metric_quartile                 vem_source
## 91                              2 Healthy Places Index Score
##    age12_plus_population age5_plus_population persons_fully_vaccinated
## 91               76365.2                82971                       18
##    persons_partially_vaccinated percent_of_population_fully_vaccinated
## 91                           22                               0.000217
##    percent_of_population_partially_vaccinated
## 91                                    0.000265
##    percent_of_population_with_1_plus_dose booster_recip_count
## 91                               0.000482                  NA
##                                                                   redacted
## 91 Information redacted in accordance with CA state privacy requirements
```

Q11. How many distinct zip codes are listed for San Diego County? -> *107 zip codes*

Q12. What San Diego County Zip code area has the largest 12 + Population in this dataset? -> *zip code 92154 has the largest 12+ population*

```
  sd_mar01<- filter(sd, as_of_date == "2022-03-01")
  mean(sd_mar01$percent_of_population_fully_vaccinated, na.rm = TRUE)
```

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

Q13. What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2022-03-01"? -> *The overall average of fully vaccinated people in San Diego are 70.5%.*

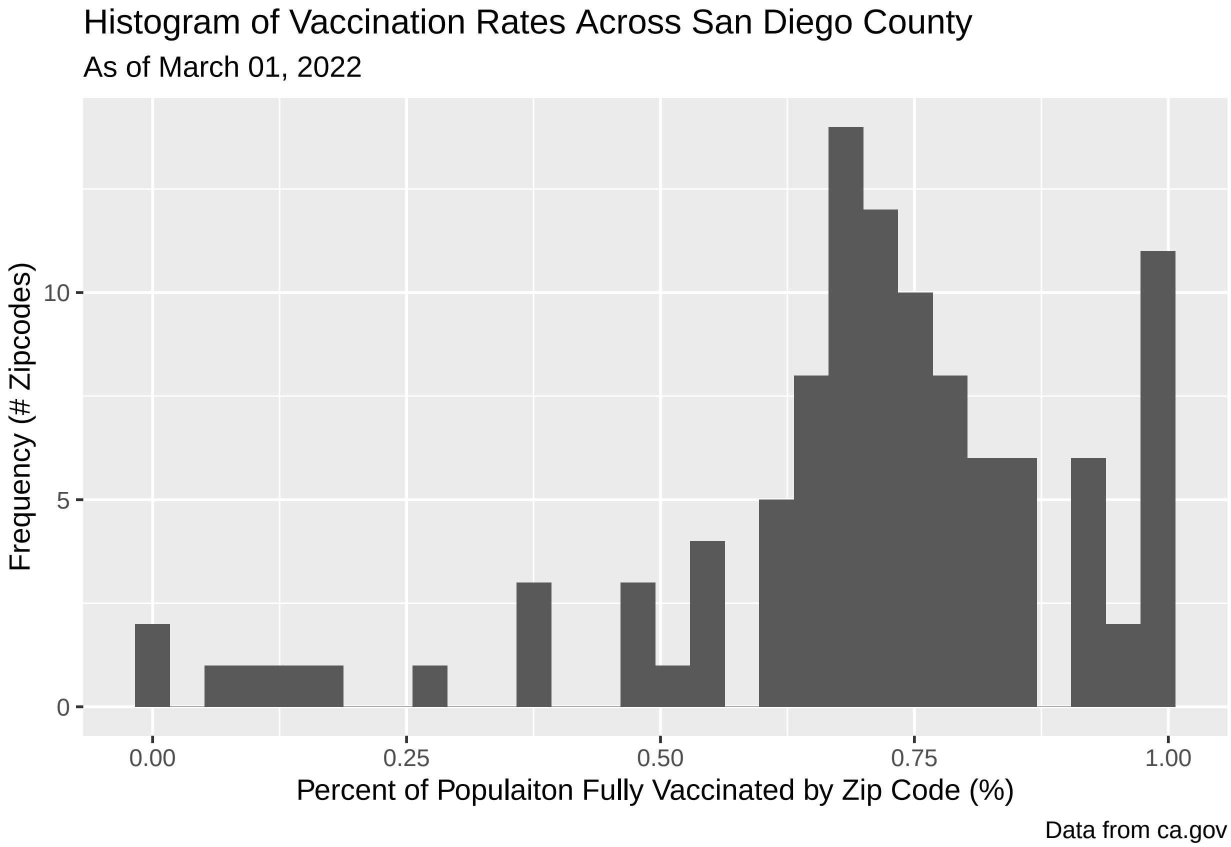
Q14. Using either ggplot or base R graphics make a summary figure that shows the distribution of Percent of Population Fully Vaccinated values as of "2022-02-22":

```
library(ggplot2)
ggplot(sd_mar01) + aes(x = sd_mar01$percent_of_population_fully_vaccinated) + geom_histogram() + labs(t
```

```
## Warning: Use of `sd_mar01$percent_of_population_fully_vaccinated` is
## discouraged. Use `percent_of_population_fully_vaccinated` instead.
```

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

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


Histogram of Vaccination Rates Across San Diego County

As of March 01, 2022

Data from ca.gov

Focus on UCSD/La Jolla

The local zip code here is 92037.

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

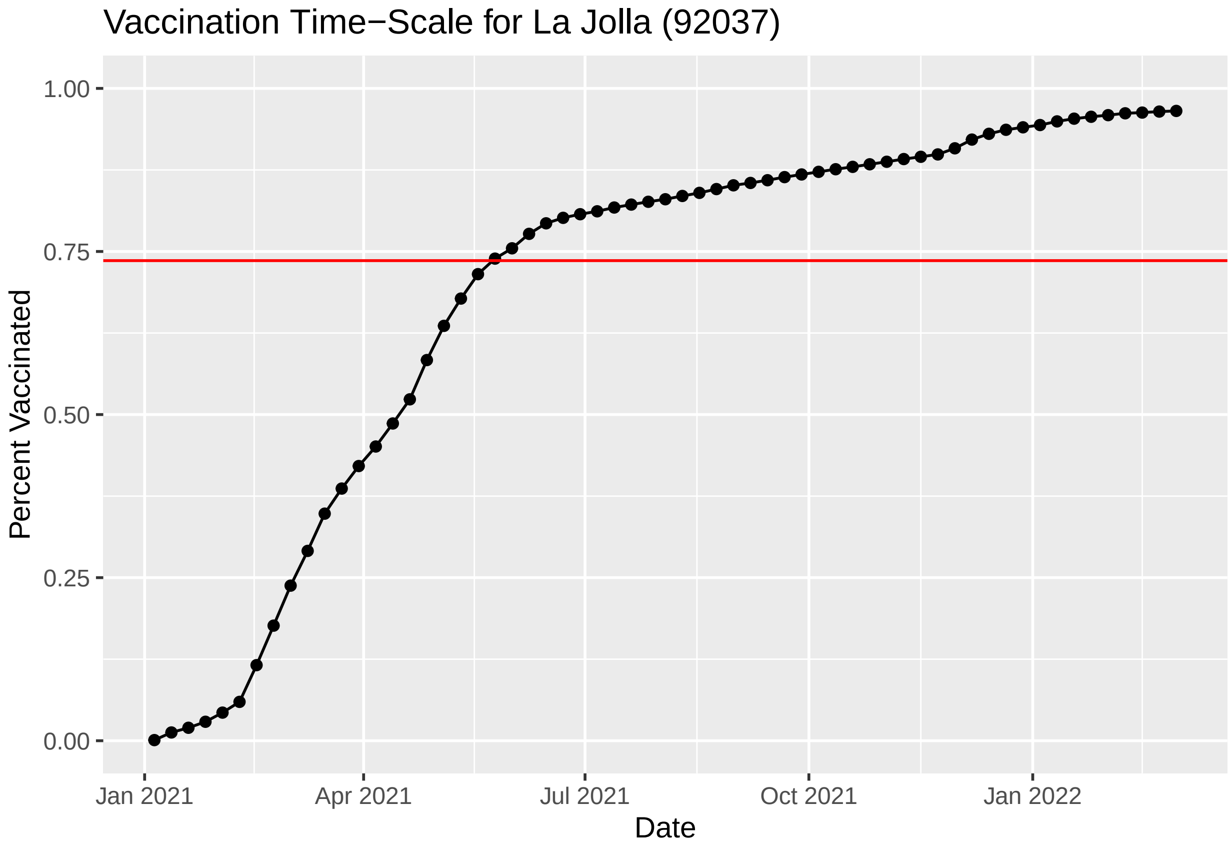
```
ucsd <- filter(sd, zip_code_tabulation_area == "92037")
ggplot(ucsd) + aes(x = ucsd$as_of_date, y = ucsd$percent_of_population_fully_vaccinated) + geom_point()
```

```
## Warning: Use of `ucsd$as_of_date` is discouraged. Use `as_of_date` instead.
```

```
## Warning: Use of `ucsd$percent_of_population_fully_vaccinated` is discouraged.
## Use `percent_of_population_fully_vaccinated` instead.
```

```
## Warning: Use of `ucsd$as_of_date` is discouraged. Use `as_of_date` instead.
```

```
## Warning: Use of `ucsd$percent_of_population_fully_vaccinated` is discouraged.
## Use `percent_of_population_fully_vaccinated` instead.
```



"This plot shows an initial slow roll out in January into February (likely due to limited vaccine availability). This is followed with rapid ramp up until a clear slowing trend from June, onward. The red line shows average rates of vaccination as of Mar 01, 2022 for similarly-sized zipcodes. Interpretation beyond this requires context from other zip code areas to answer questions such as: is this trend representative of other areas? Are more people fully vaccinated in this area compared to others?"

Comparing to similar-sized areas

Let's return to the full data set and look across every zip code area with a population at least as large as that of 92037 on as_of_date "2022-03-01".

```
ucsd[ucsd$as_of_date == "2022-03-01",]
```

```
##    as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 61 2022-03-01                    92037                 San Diego San Diego
##    vaccine_equity_metric_quartile                 vem_source
## 61                              4 Healthy Places Index Score
##    age12_plus_population age5_plus_population persons_fully_vaccinated
## 61               33675.6                36144                    34895
##    persons_partially_vaccinated percent_of_population_fully_vaccinated
## 61                        11073                               0.965444
##    percent_of_population_partially_vaccinated
## 61                                   0.306358
##    percent_of_population_with_1_plus_dose booster_recip_count redacted
## 61                                      1               16455       No
```

```
similar <- filter(vax, vax$age5_plus_population >= 36144, vax$as_of_date == "2022-03-01")
mean(similar$percent_of_population_fully_vaccinated)
```

```
## [1] 0.7359558
```

> 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 “2022-02-22”. Add this as a straight horizontal line to your plot from above with the geom_hline() function? ^^^

> 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 “2022-02-22”?

```
summary(similar$percent_of_population_fully_vaccinated)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3890  0.6554  0.7351  0.7360  0.8055  1.0000
```

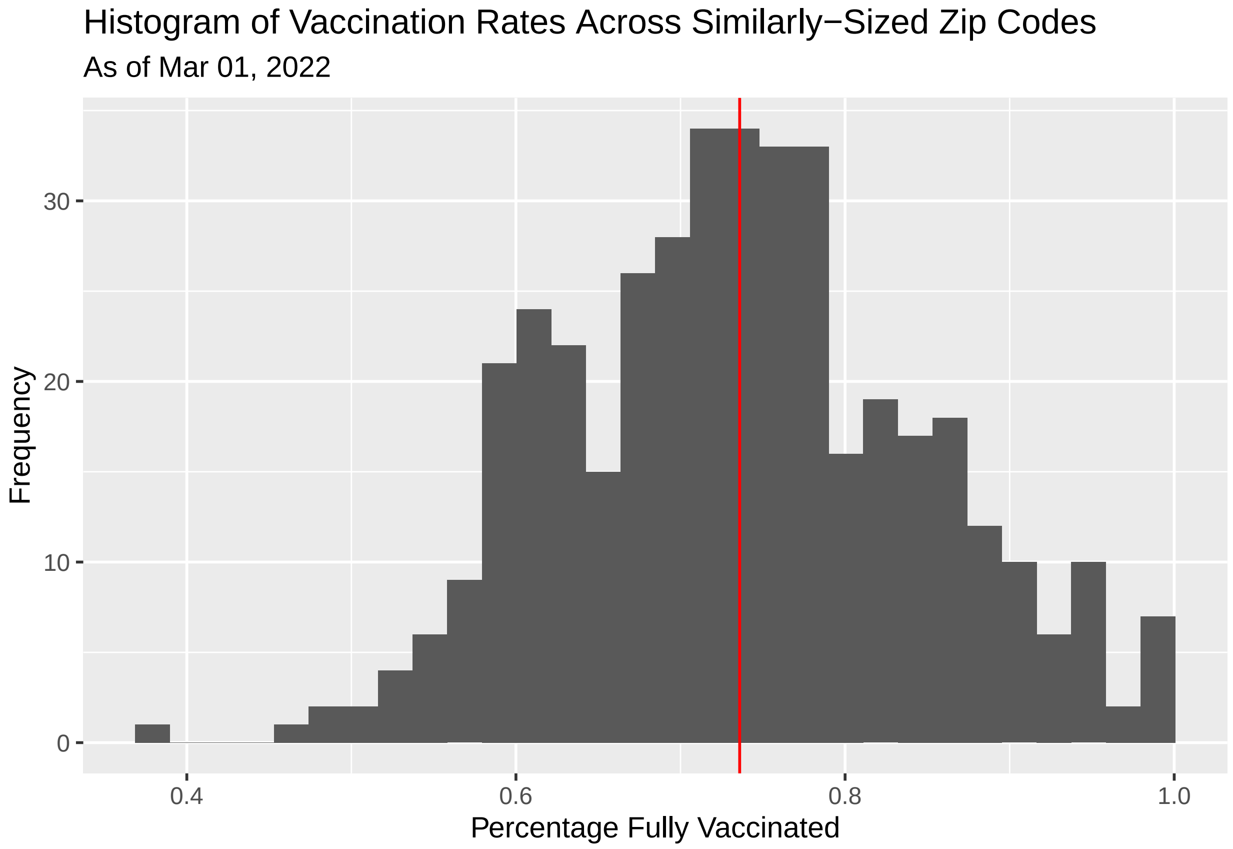
> Q18. Using ggplot generate a histogram of this data.

```
ggplot(similar) + aes(x = similar$percent_of_population_fully_vaccinated) + geom_histogram() + labs(titl
```

```
## Warning: Ignoring unknown parameters: lab
```

```
## Warning: Use of `similar$percent_of_population_fully_vaccinated` is discouraged.
## Use `percent_of_population_fully_vaccinated` instead.
```

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



Q19. Is the 92109 and 92040 ZIP code areas above or below the average value you calculated for all these above? -> *The vaccination rate for 92109 is slightly below, 92040 is significantly (significance not calculated) below*

```
vax %>% filter(as_of_date == "2022-02-22") %>%
  filter(zip_code_tabulation_area=="92109") %>%
  select(percent_of_population_fully_vaccinated) #rate = 0.723044
```

```
##   percent_of_population_fully_vaccinated
## 1                               0.723044
```

```
vax %>% filter(as_of_date == "2022-02-22") %>%
  filter(zip_code_tabulation_area=="92040") %>%
  select(percent_of_population_fully_vaccinated) #rate = 0.551304
```

```
##   percent_of_population_fully_vaccinated
## 1                               0.551304
```

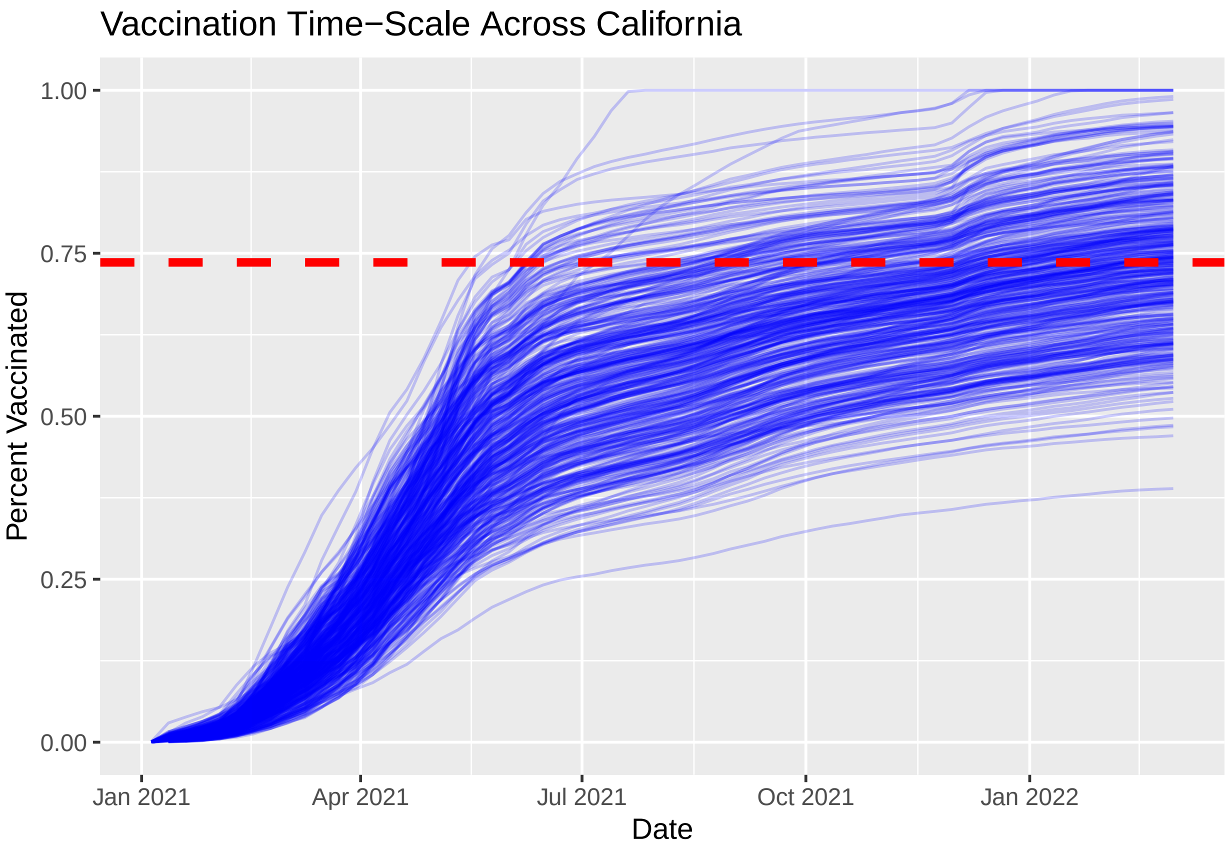
```
#Where I live :)
vax %>% filter(as_of_date == "2022-02-22") %>%
  filter(zip_code_tabulation_area=="92128") %>%
  select(percent_of_population_fully_vaccinated) #rate = 0.784705
```

```
##   percent_of_population_fully_vaccinated
## 1                               0.784705
```

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

```
similar_timecourse <- filter(vax, vax$age5_plus_population >= 36144)
ggplot(similar_timecourse) + aes(x = as_of_date, y = percent_of_population_fully_vaccinated, group = zi
```

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


Vaccination Time−Scale Across California
Percent Vaccinated
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
Date