# Project: Vaccination Rate

Rachel Diao

3/4/2022

```
vax <- read.csv('covid19vaccinesbyzipcode_test.csv')
#head(vax)
```

## Question 1

Total number of people fully vaccinated is under column "persons_fully_vaccinated".

## Question 2

Zip code tabulation area is under "zip_code_tabulation_area".

## Question 3

The earliest date is 2021-01-05.

## Question 4

The latest date is 2022-03-01.

Get an overview of the dataset with skim:

```
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 | 817.39 | 90001 | 92257.75 | 93658.50 | 95380.50 | 97635.0 | |
| vaccine_equity_metric_quartile | 307 | 0.95 | 2.44 | 1.11 | 1 | 1.00 | 2.00 | 3.00 | 4.0 | |
| age12_plus_population | 0 | 1.00 | 18895.04 | 8993.91 | 0 | 1346.95 | 13685.50 | 31756.25 | 28556.7 | |
| age5_plus_population | 0 | 1.00 | 20875.24 | 9106.02 | 0 | 1460.50 | 15364.00 | 34877.00 | 31902.0 | |
| persons_fully_vaccinated | 18338 | 0.83 | 12155.61 | 3063.88 | 11 | 1066.25 | 7374.50 | 20005.00 | 7744.0 | |
| 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 | |

### Question 5

There are 10 numeric columns in the dataset.

### Question 6

There are 18338 missing values in the persons_fully_vaccinated column

### Question 7

17.0421174% of persons_fully_vaccinated are missing.

## Working with dates

Load in package 'lubridate'!

```
library(lubridate)
```

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

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

```
today()
```

```
## [1] "2022-03-04"
```

Convert data in as_of_date column to lubridate format!

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

Now we can actually do operations on dates. To calculate difference from today - the earliest date in this dataset is

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

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

Days that the dataset spans:

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

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

How many days have I been alive? - 9025 days

```
today() - ymd('1997-06-18')
```

```
## Time difference of 9025 days
```

### Question 9

Difference between today and the last date in the dataset:3

### Question 10

There are 61 unique dates in the dataset (answer from the skimr summary)

## Working with zipcodes

Load in package zipcodeR! With geocode_zip(), we can get the centroid of the region any zipcode covers.

```
library(zipcodeR)
geocode_zip('92037')
```

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

Calculate distance (in miles) between centroids of any two zipcodes:

```
zip_distance('92037','92109')
```

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

Use reverse_zipcode() to pull lots of info on zipcodes:

```
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>
```

## Focus on San Diego county

```
# Subset to San Diego county only areas
sd <- vax[vax$county=='San Diego', ]
```

Can also do the same in 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] 6527
```

Filter for areas where population is greater than 10,000:

```
sd.10 <- filter(vax, county == "San Diego" &
                age5_plus_population > 10000)
```

### Question 11

There are 107 distinct zip codes in SD county.

### Question 12

92154 has the largest 12+ population in this dataset.

## Data for 2022-03-01

```
recent <- filter(vax, county == "San Diego", as_of_date=='2022-03-01')

#Average percent of population fully vaccinated in San Diego on this day
mean(recent$percent_of_population_fully_vaccinated, na.rm=TRUE)
```

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

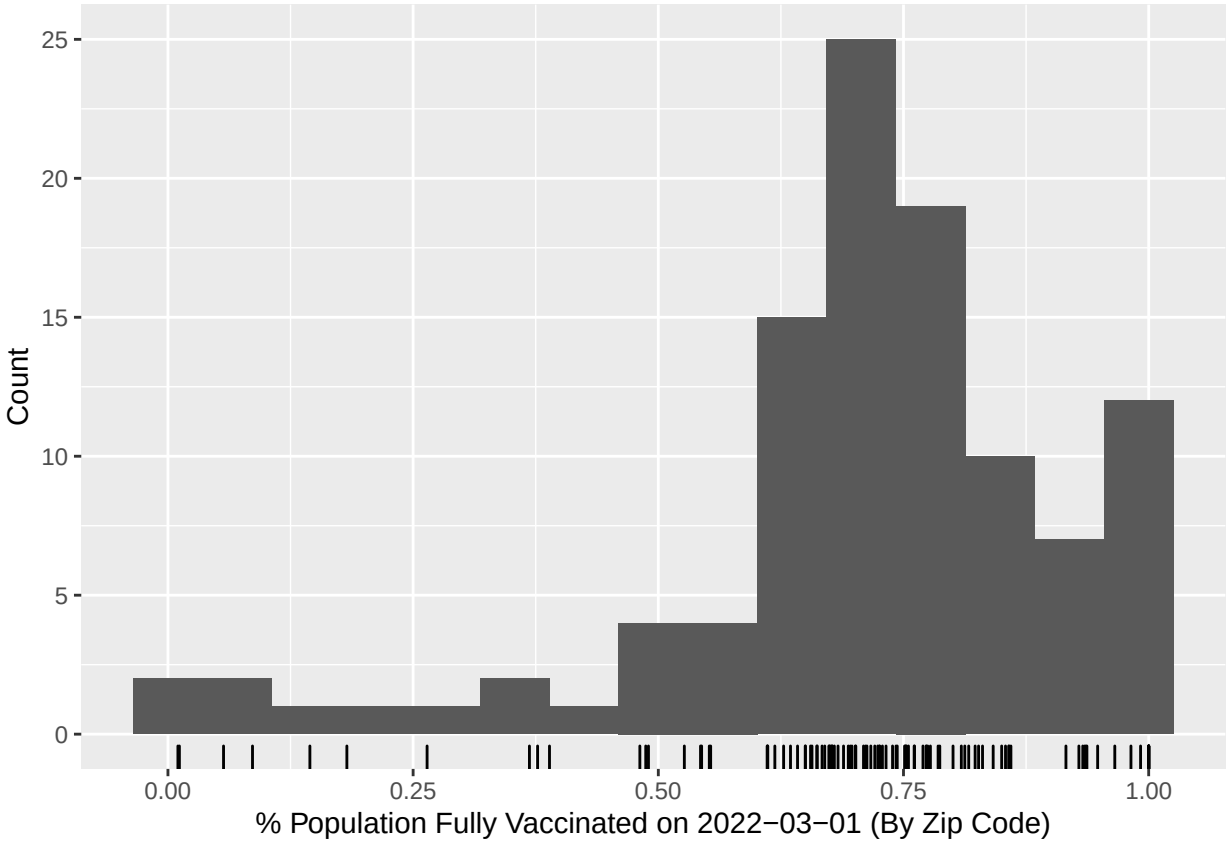
### Question 13

70.53% of the population in San Diego was vaccinated by 03-01-2022.

### Question 14

```
library(ggplot2)
ggplot(recent, aes(x=percent_of_population_fully_vaccinated)) +
  geom_histogram(bins=15) + geom_rug() +
  xlab('% Population Fully Vaccinated on 2022-03-01 (By Zip Code)') +
  ylab('Count')
```

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




# UCSD data

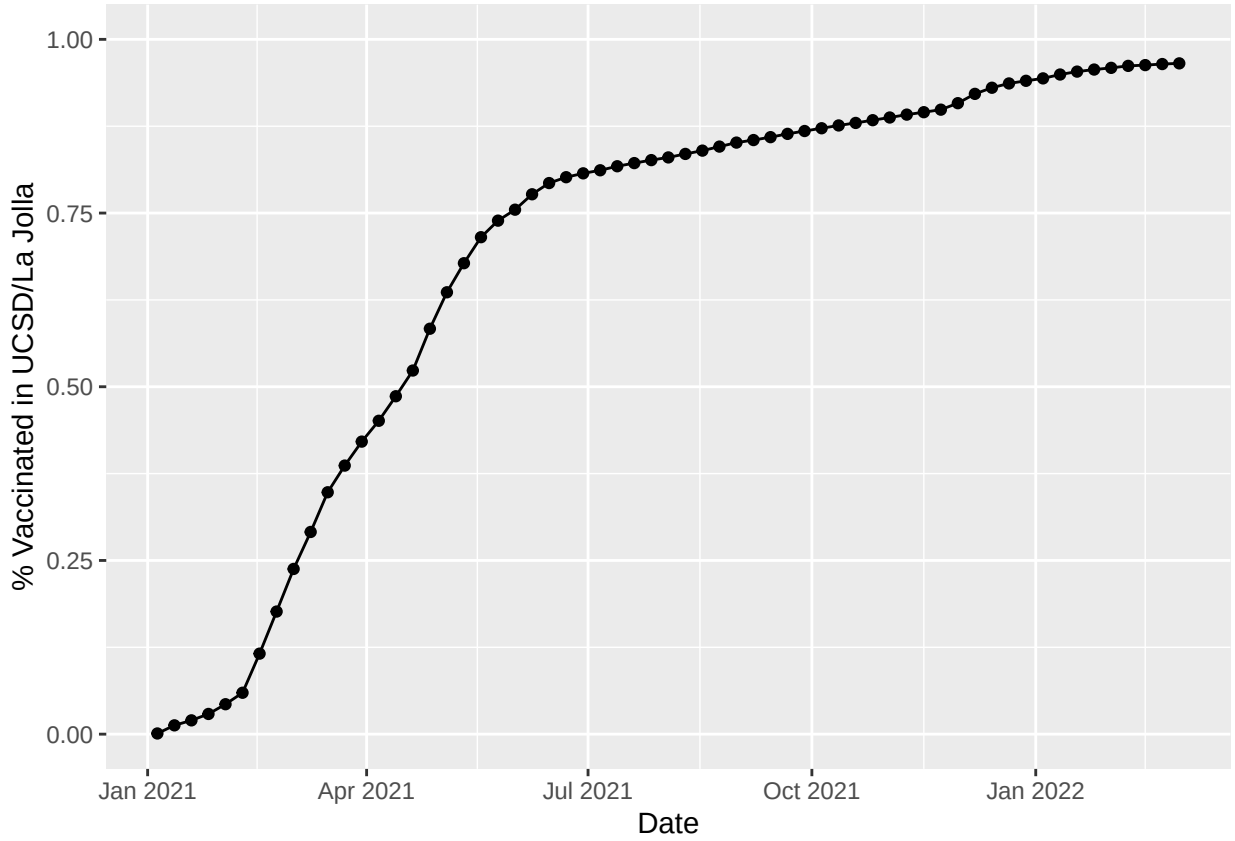
```
ucsd <- filter(sd, zip_code_tabulation_area=="92037")
ucsd[1,]$age5_plus_population
```

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

### Question 15

Vaccination rate time-course for UCSD zip code:

```
ggplot(ucsd) +
  aes(x=as_of_date,
      y=percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) +
  ylim(c(0,1)) +
  labs(x='Date', y="% Vaccinated in UCSD/La Jolla")
```



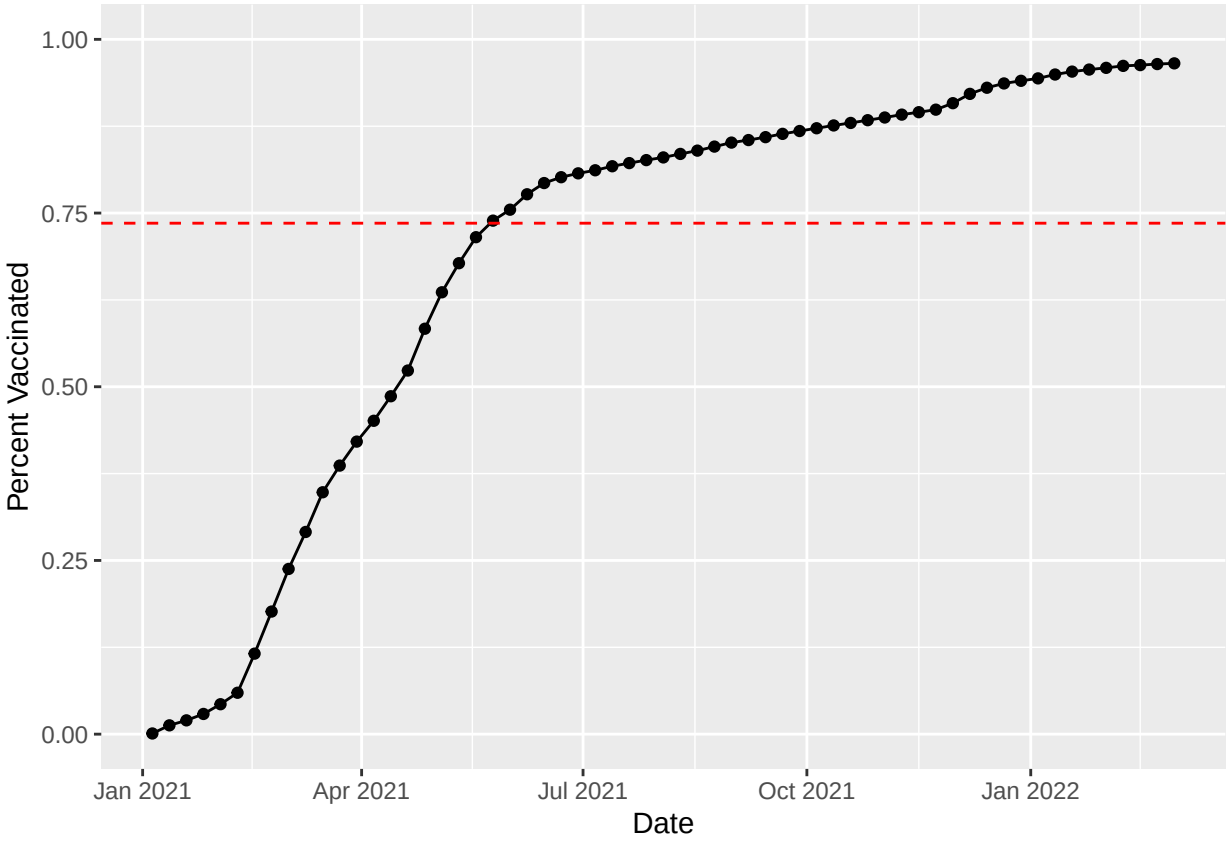

# Compare to similarly-sized areas:

```
# Subset to all CA areas with a population as large as 92037
vax.36 <- filter(vax, age5_plus_population > 36144 &
                 as_of_date == "2022-03-01")
```

## Question 16

73.5397433% people are vaccinated in areas as large as 92037.

```
ggplot(ucsd) +
  aes(x=as_of_date,
      y=percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) +
  geom_hline(yintercept=mean(vax.36$percent_of_population_fully_vaccinated),
             linetype='dashed', col='red') +
  ylim(c(0,1)) +
  labs(x='Date', y="Percent Vaccinated")
```




## Question 17

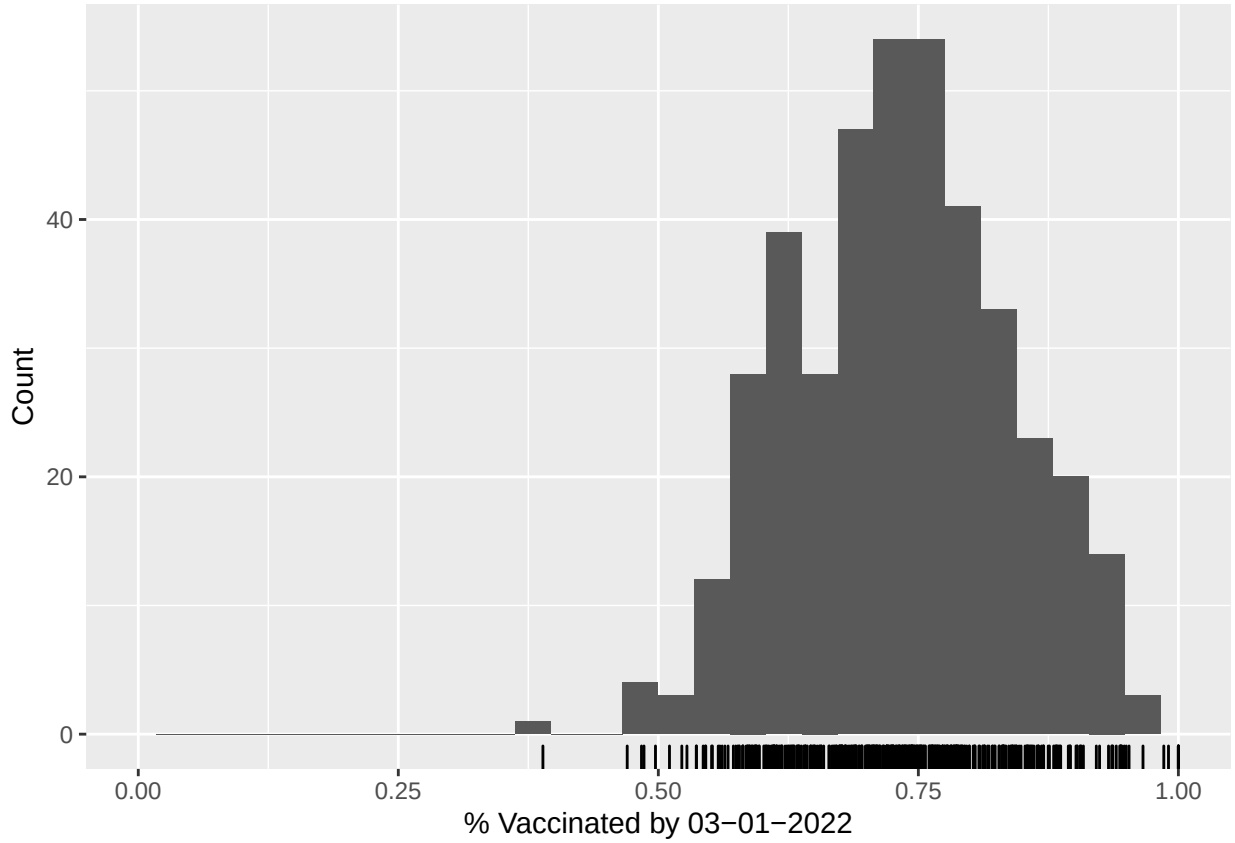
```
summary(vax.36$percent_of_population_fully_vaccinated)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3890  0.6554  0.7350  0.7354  0.8044  1.0000
```

## Question 18

```
ggplot(vax.36) + aes(x=percent_of_population_fully_vaccinated) +
  geom_histogram(bins=30) + geom_rug() +
  xlim(c(0,1)) +
  xlab('% Vaccinated by 03-01-2022') + ylab('Count')
```

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




## Question 19

92109 (55.20%) and 92040 (72.38%) averages are both below the average % vaccinated for all counties in California with population size similar or larger than 92037, though 92109 is only about 1% lower than the average.

```
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
```

```
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
```

## 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(c(0,1)) +
  labs(x='Date', y='% of Population Vaccinated (by Zip Code)',
       title='Vaccination Rates Across California',
       subtitle='Only areas with population above 36000 are shown') +
  geom_hline(yintercept = mean(vax.36$percent_of_population_fully_vaccinated), linetype='dashed')
```

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