COVID-19 Vaccination Project

Yash Garodia

02/03/2022

The goal of this hands-on mini-project is to examine and compare the Covid-19 vaccination rates around San Diego.

We will start by downloading the most recently dated "Statewide COVID-19 Vaccines Administered by ZIP Code" CSV file from: https://data.ca.gov/dataset/covid-19-vaccine-progress-dashboard-data-by-zip-code

Whilst you are on this website have a look at the Data Dictionary file that explains the various columns within the CSV file that you just downloaded.

There are also important notes about the limitations of the data. For example: "These data do NOT include doses administered by the following federal agencies who received vaccine allocated directly from CDC: Indian Health Service, Veterans Health Administration, Department of Defense, and the Federal Bureau of Prisons." One obvious implication here would be that Zip code areas that include military bases will likely show artificially low vaccination rates. We will bare this in mind for later.

Getting Started

Be sure to move your downloaded CSV file to your project directory and then read/import into an R object called vax. We will use this data to answer all the questions below.

```
#Import vaccination data

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

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction        county
## 1 2021-01-05                   92140                 San Diego     San Diego
## 2 2021-01-05                   94133             San Francisco San Francisco
## 3 2021-01-05                   94523              Contra Costa  Contra Costa
## 4 2021-01-05                   94005                 San Mateo     San Mateo
## 5 2021-01-05                   94104             San Francisco San Francisco
## 6 2021-01-05                   94549              Contra Costa  Contra Costa
##   vaccine_equity_metric_quartile               vem_source
## 1                             NA            No VEM Assigned
## 2                              3 Healthy Places Index Score
## 3                              4 Healthy Places Index Score
## 4                              4 Healthy Places Index Score
## 5                             NA            No VEM Assigned
## 6                              4 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1                3747.7                3737                       NA
```

```
## 2               25070.5               25957                       NA
## 3               30457.9               32828                       NA
## 4                3996.1                4364                       NA
## 5                 387.8                 399                       NA
## 6               25393.8               28468                       NA
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                           NA                                     NA
## 2                           NA                                     NA
## 3                           NA                                     NA
## 4                           NA                                     NA
## 5                           NA                                     NA
## 6                           NA                                     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
##                                                        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?

To find the earliest date, we can use the min() function.

```
min(vax$as_of_date)
```

```
## [1] "2021-01-05"
```

The earliest date is 2021-01-05

Q4. What is the latest date in this dataset?

To find the latest date, we can use the max() function.

```
max(vax$as_of_date)
```

```
## [1] "2022-02-22"
```

The latest date is 2022-02-22

As we have done previously, let's call the skim() function from the skimr package to get a quick overview of this dataset:

```
library(skimr)
```

```
## Warning: package 'skimr' was built under R version 4.0.2
```

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

Table 1: Data summary

Name	vax
Number of rows	105840
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	60	0
local_health_jurisdiction	0	1	0	15	300	62	0
county	0	1	0	15	300	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	5220	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	
age12_plus_population	0	1.00	18895.04	18993.92	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	18174	0.83	12064.29	12983.91	11	1059.00	7287.50	19859.00	77213.0	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
persons_partially_vaccinated	18174	0.83	820.71	1318.77	11	76.00	370.00	1066.00	31869.0	
percent_of_population_fully_vaccinated	18174	0.83	0.51	0.26	0	0.33	0.54	0.70	1.0	
percent_of_population_partially_vaccinated	18174	0.83	0.05	0.09	0	0.01	0.03	0.05	1.0	
percent_of_population_with_plus_dose	18174	0.83	0.54	0.27	0	0.35	0.58	0.75	1.0	
booster_recip_count	64191	0.39	3923.43	5704.10	11	169.00	1072.00	5803.00	49951.0	

Q5. How many numeric columns are in this dataset?

There are 10 numeric columns in this dataset

Q6. Note that there are "missing values" in the dataset. How many NA values there in the persons_fully_vaccinated column?

There are 18174 NA values in the persons_fully_vaccinated column, as shown in the n_missing section of skim(vax)

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

```
#Percentage missing = Missing/Total
#To find total missing
nrow(vax)
```

```
## [1] 105840
```

```
round(sum( is.na(vax$persons_fully_vaccinated) )/nrow(vax),2)*100
```

```
## [1] 17
```

17% of persons_fully_vaccinated values are missing.

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

This data may be missing because certain areas in san diego may not have reported their data as of yet, or are still taking readings.

Working with dates

One of the "character" columns of the data is as_of_date, which contains dates in the Year-Month-Day format.

Dates and times can be annoying to work with at the best of times. However, in R we have the excellent lubridate package, which can make life allot easier. Here is a quick example to get you started:

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.0.2
```

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

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

What is today's date (at the time I am writing this obviously)

```
today()
```

```
## [1] "2022-03-02"
```

The as_of_date column of our data is currently not that usable. For example we can't easily do math with it like answering the simple question how many days have passed since data was first recorded:

```
# This will give an Error!
#today() - vax$as_of_date[1]
```

However if we convert our date data into a lubridate format things like this will be much easier as well as plotting time series data later on.

```
# Specify that we are using the year-month-day format
vax$as_of_date <- ymd(vax$as_of_date)
```

Now we can do math with dates. For example: How many days have passed since the first vaccination reported in this dataset?

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

```
## Time difference of 421 days
```

Using the last and the first date value we can now determine how many days the dataset span?

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

```
## Time difference of 413 days
```

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

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

```
## Time difference of 8 days
```

It has been 8 days since the last update of the dataset.

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

In order to do this, we must find the unique dates and then use the length() function to find the total number of them.

```
dates <- c(vax$as_of_date)
length(unique(dates))
```

```
## [1] 60
```

There are 60 different dates.

Working with ZIP codes

One of the numeric columns in the dataset (namely vax$zip_code_tabulation_area) are actually ZIP codes - a postal code used by the United States Postal Service (USPS). In R we can use the zipcodeR package to make working with these codes easier. For example, let's install and then load up this package and to find the centroid of the La Jolla 92037 (i.e. UC San Diego) ZIP code area.

```
library(zipcodeR)
```

```
## Warning: package 'zipcodeR' was built under R version 4.0.2
```

```
geocode_zip(92037)
```

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

Calculate the distance between the centroids of any two ZIP codes in miles, e.g.

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

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

More usefully, we can pull census data about ZIP code areas (including median household income etc.). For example:

```
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>
```

Optional: We can use this reverse_zipcode() to pull census data later on for any or all ZIP code areas we might be interested in.

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

We could also access socioeconomic data for different ZIP code areas in a similar way if we wanted to investigate factors that might be correlated with different vaccine uptake rates.

Another informative data exploration might be to plot the various values along with the ZIP codes latitude and longitude values on a map using a package like leafelet or using ggplot2 itself similar to this post. For now we will leave this as an optional extension exercise.

Focus on the San Diego Area

Let's now focus in on the San Diego County area by restricting ourselves first to vax$county == "San Diego" entries. We have two main choices on how to do this. The first using base R the second using the dplyr package:

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

Using dplyr the code would look like this:

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
##
## 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] 6420
```

Using dplyr is often more convenient when we are subsetting across multiple criteria - for example all San Diego county areas with a population of over 10,000.

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

Q11. How many distinct zip codes are listed for San Diego County?

All zip code tabulation data is stored in the zip_code_tablulation_area column.

```
sd.zipcodes <- sd$zip_code_tabulation_area
length(unique(sd.zipcodes))
```

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

There are a total of 107 distinct zip codes listed for San Diego county.

Q12. What San Diego County Zip code area has the largest 12 + Population in this dataset?

First we must use the dplyr package to sort the sd dataset in descending order of 12 + population, and then return the zipcode for the first.

```
sd.12 <- sd %>% as.data.frame() %>% arrange(desc(age12_plus_population))
sd.12[1,2]
```

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

Zipcode area 92154 has the largest 12 + Population in this dataset.

Q13. What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2022-02-22"?

First we must filter the sd dataset to contain only those results that have as_of_date as 2022-02-22, then we store only the 'percent_of_population_fully_vaccinated' data in a separate variable and calculate the mean, removing NA values from consideration.

```
sd.2022 <- filter(sd, as_of_date == "2022-02-22")
sd2022022.vaccinated <- c(sd.2022$percent_of_population_fully_vaccinated)
mean(sd2022022.vaccinated, na.rm = TRUE)
```

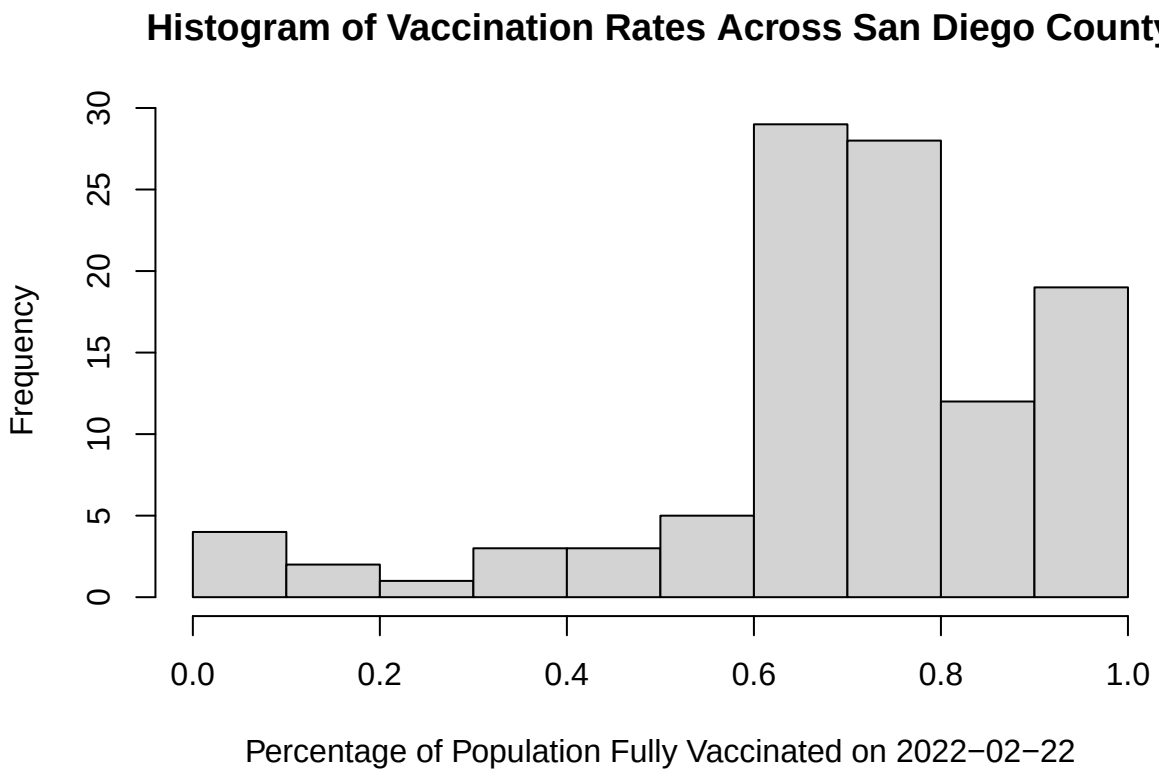
```
## [1] 0.7030846
```

The overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2022-02-22" is 70.3%

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"?

We can use the base R hist function for this, providing x labels and the main title:

```
hist(sd.2022$percent_of_population_fully_vaccinated, xlab = "Percentage of Population Fully Vaccinated
```



Focus on UCSD/La Jolla

UC San Diego resides in the 92037 ZIP code area and is listed with an age 5+ population size of 36,144.

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

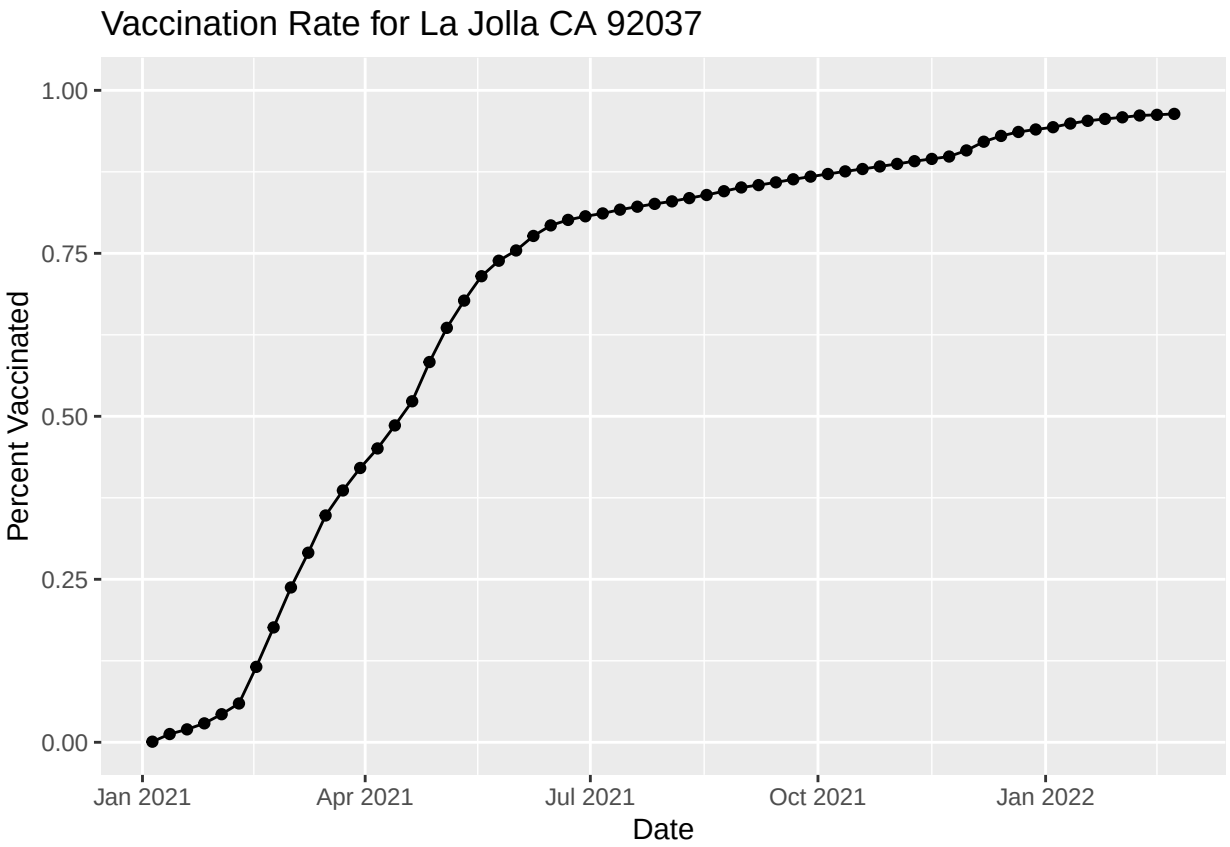
```
## [1] 36144
```

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

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

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



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 "2022-02-22".

```
# Subset to all CA areas with a population as large as 92037
vax.36 <- filter(vax, age5_plus_population > 36144 &
                 as_of_date == "2022-02-22")
head(vax.36)
```

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction       county
## 1 2022-02-22                   94582             Contra Costa Contra Costa
## 2 2022-02-22                   92592                Riverside    Riverside
## 3 2022-02-22                   92504                Riverside    Riverside
## 4 2022-02-22                   94546                  Alameda      Alameda
## 5 2022-02-22                   94577                  Alameda      Alameda
## 6 2022-02-22                   94565             Contra Costa Contra Costa
##   vaccine_equity_metric_quartile                vem_source
## 1                              4 Healthy Places Index Score
## 2                              3 Healthy Places Index Score
## 3                              2 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               34809.5               40433                    42744
## 2               69581.7               79782                    44648
## 3               50996.7               56235                    32781
## 4               37839.8               41600                    37452
## 5               42041.7               45192                    39770
## 6               80663.4               90579                    74795
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                         2755                                1.000000
## 2                         5809                                0.559625
## 3                         3205                                0.582929
## 4                         3070                                0.900288
## 5                         2529                                0.880023
## 6                         5135                                0.825743
##   percent_of_population_partially_vaccinated
## 1                                   0.068137
## 2                                   0.072811
## 3                                   0.056993
## 4                                   0.073798
## 5                                   0.055961
## 6                                   0.056691
##   percent_of_population_with_1_plus_dose booster_recip_count redacted
## 1                               1.000000               27798       No
## 2                               0.632436               20599       No
## 3                               0.639922               14119       No
## 4                               0.974086               23191       No
## 5                               0.935984               24164       No
## 6                               0.882434               36596       No
```

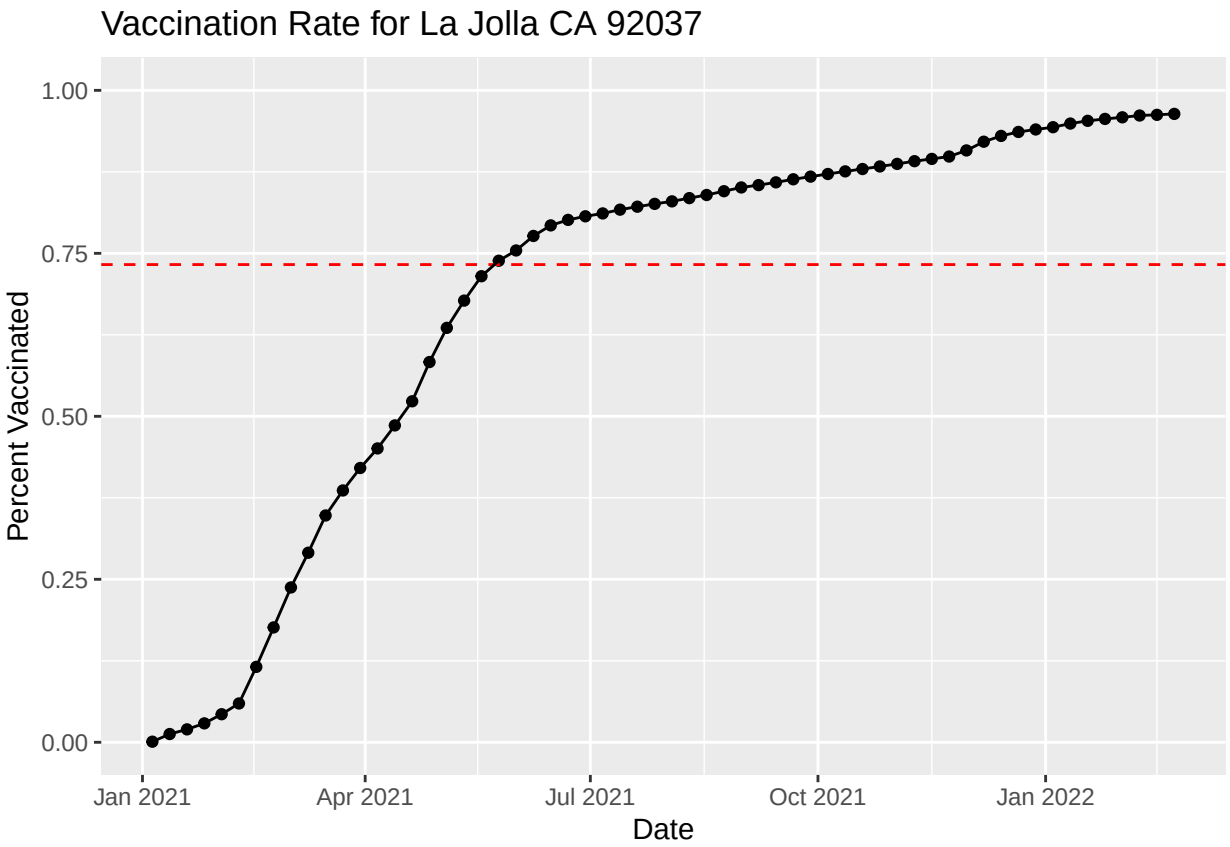
Q16. Calculate the mean "Percent of Population Fully Vaccinated" for ZIP code areas with a population as large as 92037 (La Jolla) as_of_date "2022-02-22". Add this as a straight horizontal line to your plot from above with the geom_hline() function?

We can store the "Percent of Population Fully Vaccinated" data of vax.36 in a separate variable and calculate the mean:

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

Now to add this to the earlier plot, we can use the geom_hline() function:

```
ggplot(ucsd) + aes(as_of_date, percent_of_population_fully_vaccinated) + geom_point() + geom_hline(yint
```



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"?

We can use the summary function for this.

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3878  0.6534  0.7327  0.7327  0.8024  1.0000
```

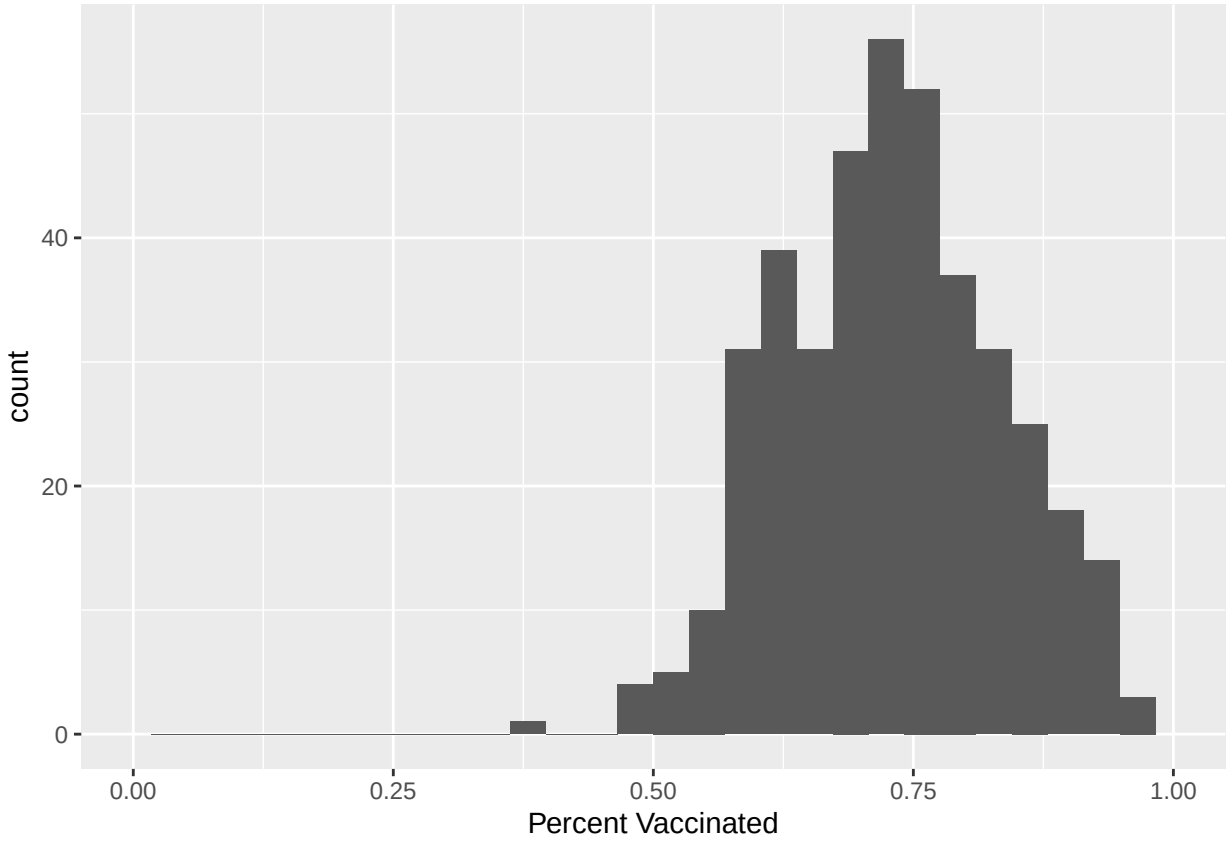
Min : 38.78%, 1st Quartile : 65.34%, Median : 73.27%, Mean : 73.27%, 3rd Quartile : 80.24%, Max = 100%

Q18. Using ggplot generate a histogram of this data.

```
ggplot(vax.36) + aes(percent_of_population_fully_vaccinated) + geom_histogram() + labs(x = "Percent Vac
```

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

```
##   percent_of_population_fully_vaccinated
## 1                                0.55093
```

```
summary(vax$percent_of_population_fully_vaccinated)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.327   0.537   0.506   0.698   1.000   18174
```

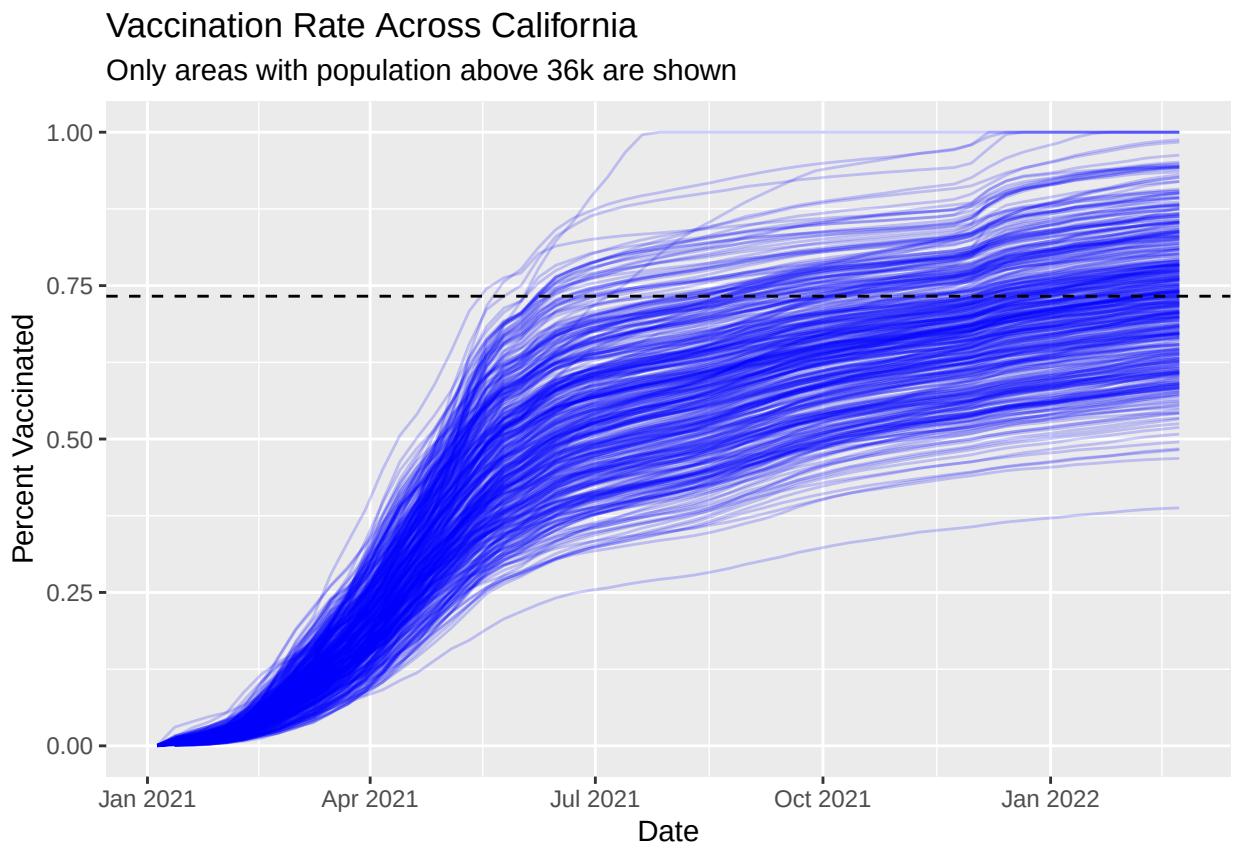
The 92109 and 92040 ZIP code areas have lower average values than 92037.

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(c(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 = ppfv_mean, linetype= 2)
```

```
## Warning: Removed 309 row(s) containing missing values (geom_path).
```



Q21. How do you feel about traveling for Spring Break and meeting for in-person class afterwards?

Super excited! Can't wait :)