

class14_Covid_Mini-Project

Seong Tae Gwon

3/7/2022

Background

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>

Getting Started

```
# 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           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
```

Q1. What column details the total number of people fully vaccinated?

Answer: The column `persons_fully_vaccinated` details the total number of people fully vaccinated.

Q2. What column details the Zip code tabulation area?

Answer: The column `zip_code_tabulation_area` details the Zip code tabulation area.

Q3. What is the earliest date in this dataset?

```
head(vax$as_of_date)
```

```
## [1] "2021-01-05" "2021-01-05" "2021-01-05" "2021-01-05" "2021-01-05"
## [6] "2021-01-05"
```

Answer: The earliest date in this dataset is 2021-01-05 (YYYY-MM-DD).

Q4. What is the latest date in this dataset?

```
tail(vax$as_of_date)
```

```
## [1] "2022-03-01" "2022-03-01" "2022-03-01" "2022-03-01" "2022-03-01"
## [6] "2022-03-01"
```

Answer: As of 3/1/22, the latest date in this dataset is 2022-03-01 (YYYY-MM-DD).
 Note: Lab has 2022-02-22 as the latest date.

Let's call the skim() function from the skimr package to get a quick overview of this dataset.

```
#install.packages("devtools")
#devtools::install_github("ropensci/skimr")
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	1817.39	90001	92257.75	93658.50	95380.50	97635.0	
vaccine_equity_metric_quarter	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	1756.12	88556.7	
age5_plus_population	0	1.00	20875.24	1106.02	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_plus_doses	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	

Q5. How many numeric columns are in this dataset?

Answer: As of 3/1/22, there are 10 numeric columns in this dataset.

Note: Lab has 9 numeric columns as answer.

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

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

```
## [1] 18338
```

Answer: As of 3/1/22, there are 18,338 NA values in the `persons_fully_vaccinated` column.

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

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

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

Answer: As of 3/1/22, 17.04% of `persons_fully_vaccinated` values are missing.

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

Answer: Individuals currently in US vaccine clinical trials are considered unvaccinated until the trial has concluded (2 weeks).

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 a lot easier.

```
#install.packages("lubridate")
library(lubridate)
```

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

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

What is today’s date?

```
today()
```

```
## [1] "2022-03-07"
```

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 of how many days have passed since data was first recorded.

However, if we convert our date data into a `lubridate` format, 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 426 days
```

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

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

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

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

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

```
## Time difference of 6 days
```

Answer: It has been 6 days since the last update of this dataset.

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

```
length(unique(vax$as_of_date))
```

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

Answer: There are 61 unique dates in this dataset.

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.

```
#install.packages("zipcodeR")
library(zipcodeR)
# Centroid of LA Jolla 92037 ZIP code area
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>
```

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)
```

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 code
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
```

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?

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

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

Answer: There are 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?

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

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

Answer: ZIP code 92154 has the largest 12+ population.

Using dplyr select all San Diego “county” entries on “as_of_date” “2022-02-22” and use this for the following questions.

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

```
sd.percentfc <- filter(vax, county == "San Diego" & as_of_date == "2022-02-22")  
mean(sd.percentfc$percent_of_population_fully_vaccinated, na.rm = TRUE)
```

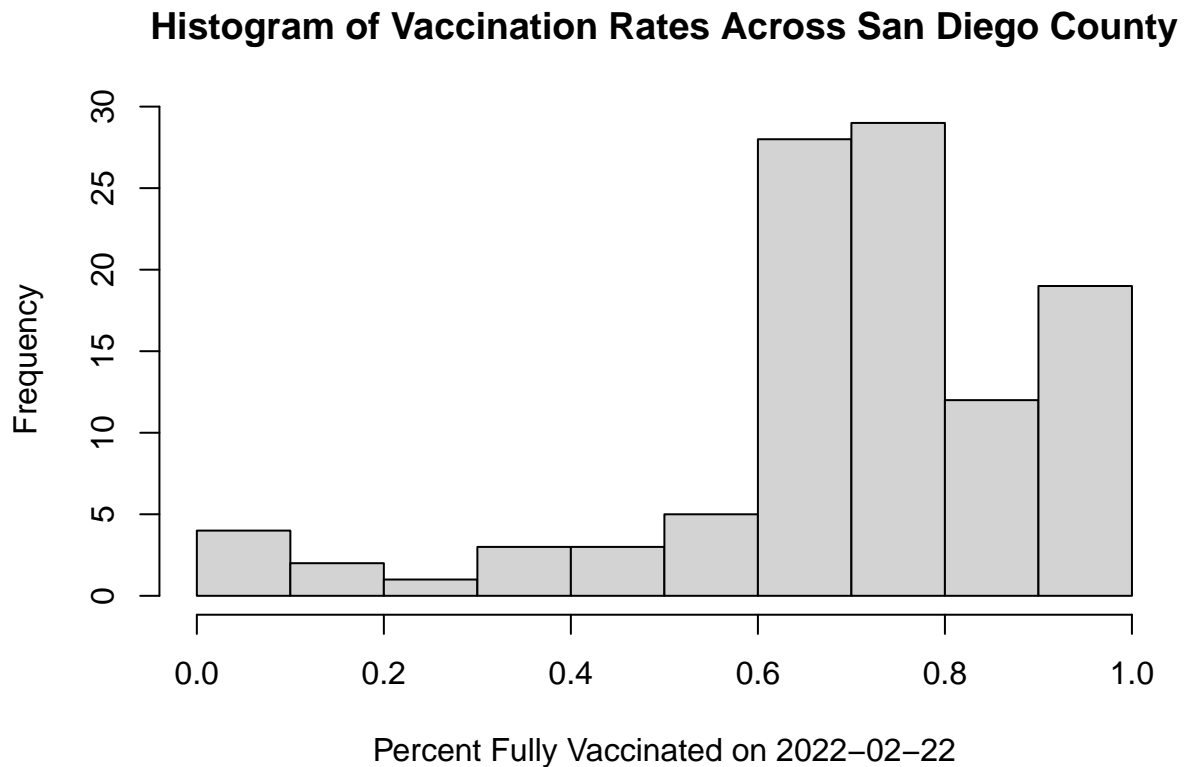
```
## [1] 0.7041551
```

Answer: The overall average “Percent of Population Fully Vaccinated” value is 0.7042.

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)

hist(sd.percentfc$percent_of_population_fully_vaccinated,
     main="Histogram of Vaccination Rates Across San Diego County",
     xlab="Percent Fully Vaccinated on 2022-02-22",
     ylab="Frequency")
```

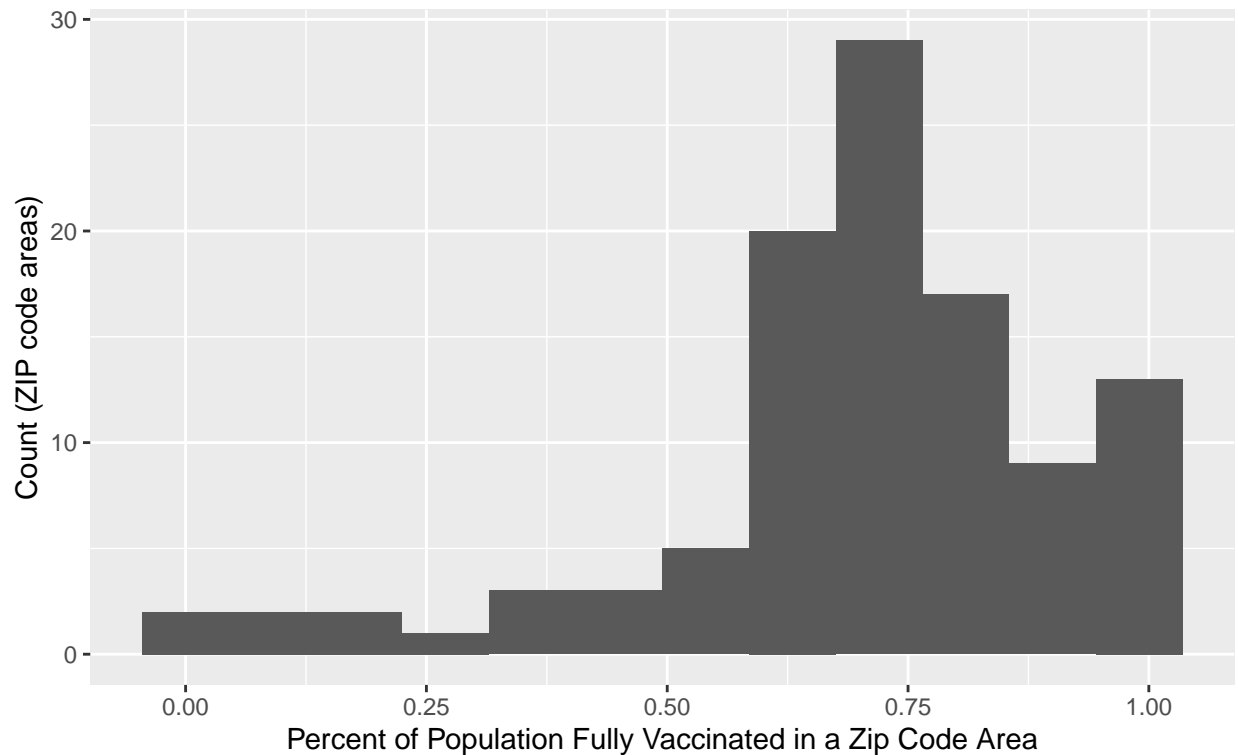


Answer:

```
ggplot(sd.percentfc) +
  aes(percent_of_population_fully_vaccinated) +
  geom_histogram(bins = 12) +
  labs(x="Percent of Population Fully Vaccinated in a Zip Code Area",
       y="Count (ZIP code areas)",
       title="Histogram of Vaccination Rates Across San Diego County",
       subtitle="As of 2022-02-22")
```

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


Histogram of Vaccination Rates Across San Diego County
As of 2022-02-22



Focus on UCSD/La Jolla

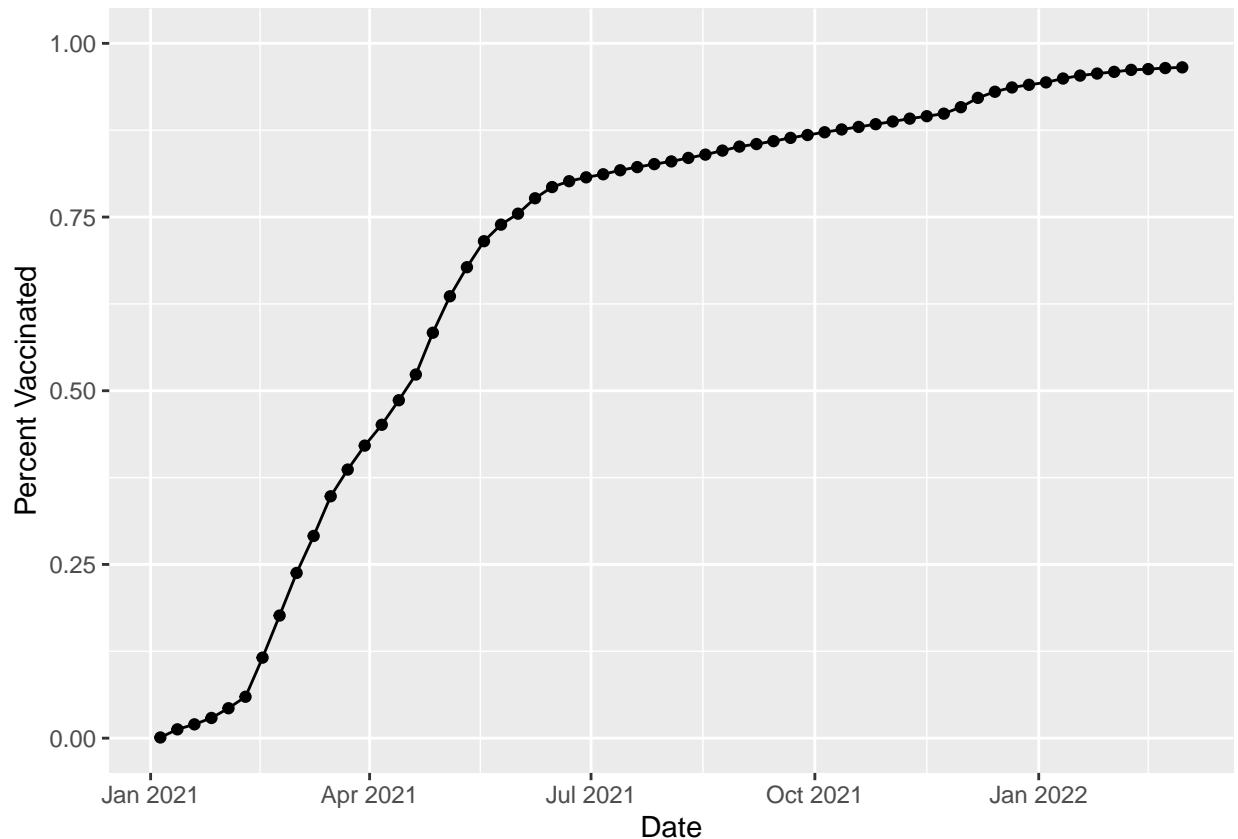
UC San Diego resides in the 92037 ZIP code area and is listed with an age 5+ population size of 36,144.

```
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.

```
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")
```



Answer:

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)
```

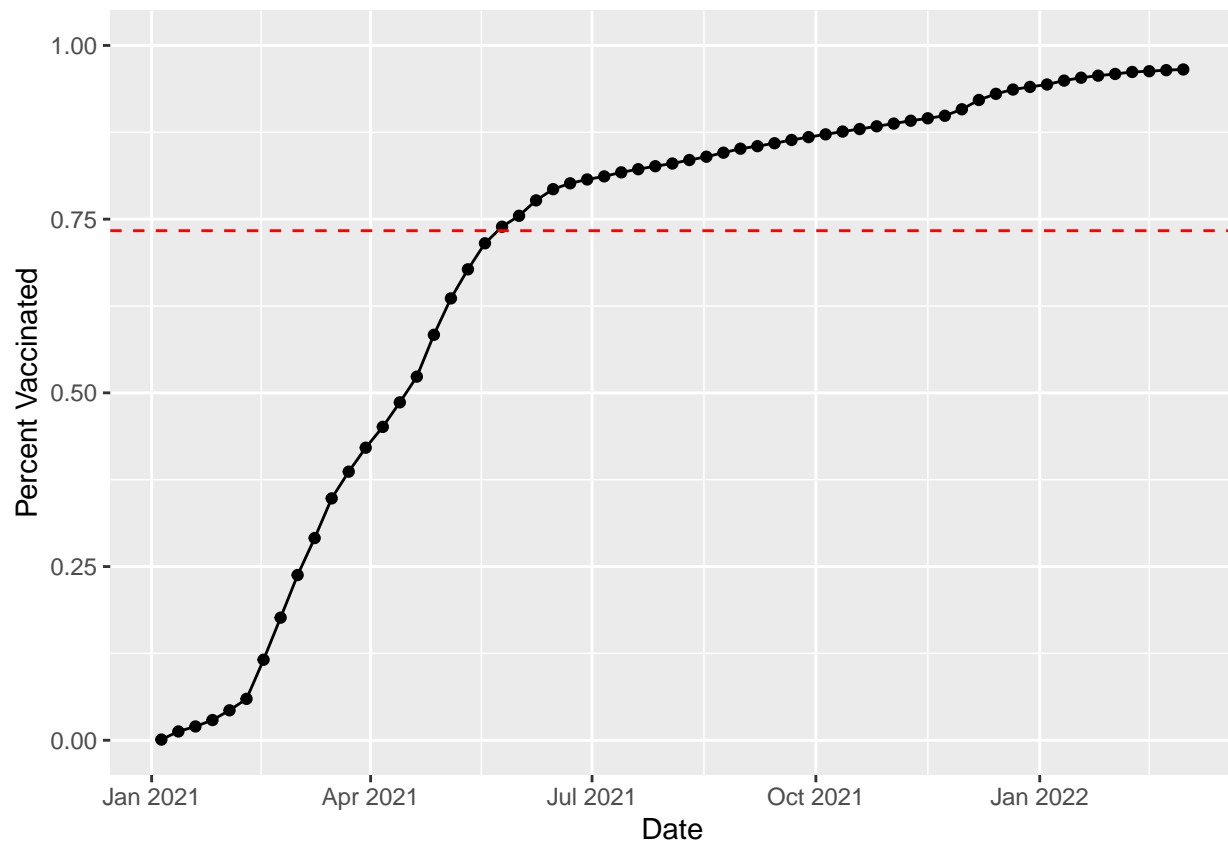
Q16. Calculate the mean "Percent of Population Fully Vaccinated" for ZIP code areas with a population as large as 92037 (La Jolla) as_of_date "2022-02-22". Add this as a straight horizontal line to your plot from above with the geom_hline() function?

```
mean.36 <- mean(vax.36$percent_of_population_fully_vaccinated)
mean.36
```

```
## [1] 0.733385
```

Answer: 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" is 0.7334.

```
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") +
  geom_hline(yintercept=mean.36, col = "red", linetype = "dashed")
```



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”?

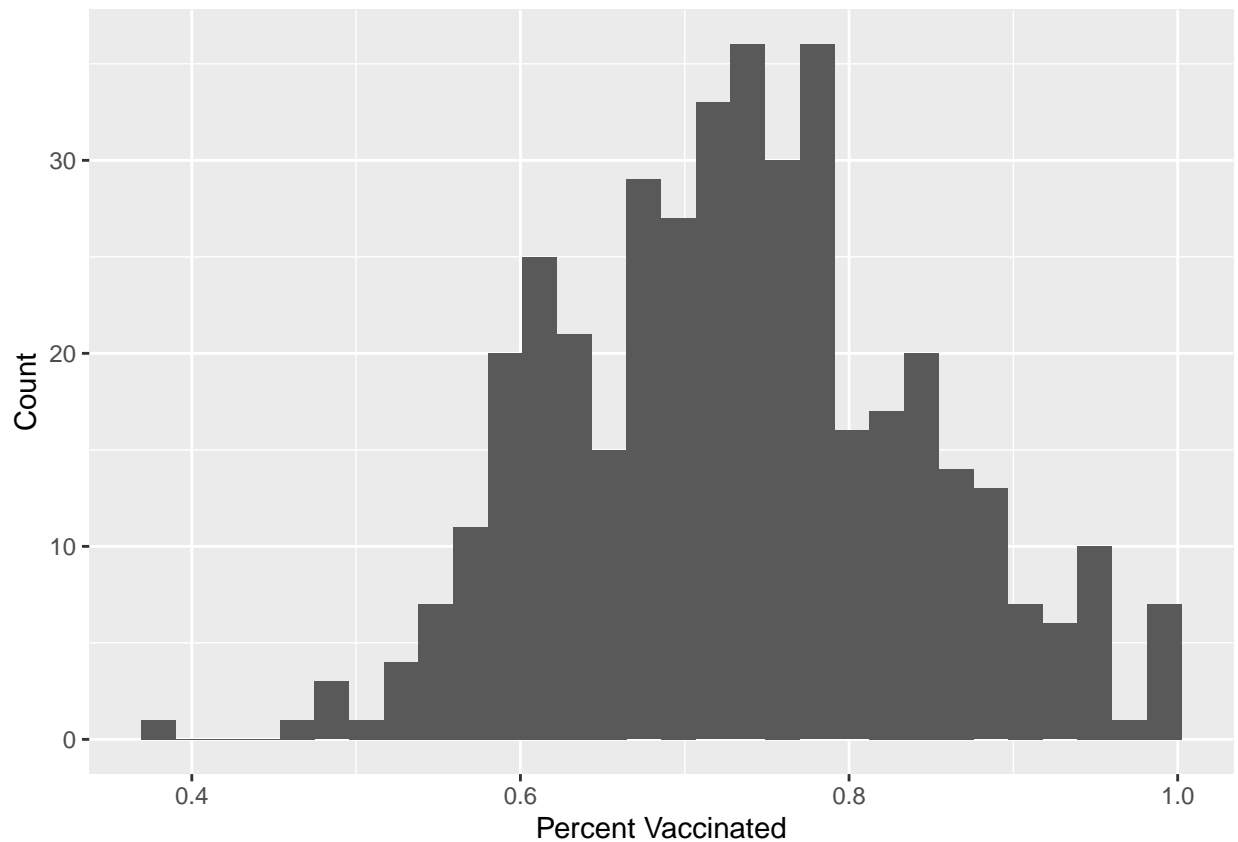
```
summary(vax.36$percent_of_population_fully_vaccinated)
```

Answer:

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.3881  0.6539   0.7333   0.7334  0.8027   1.0000
```

Q18. Using ggplot generate a histogram of this data.

```
ggplot(vax.36)+
  aes(percent_of_population_fully_vaccinated) +
  geom_histogram(bins=30) +
  labs(x="Percent Vaccinated", y="Count")
```



Answer:

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 0.551304
```

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

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

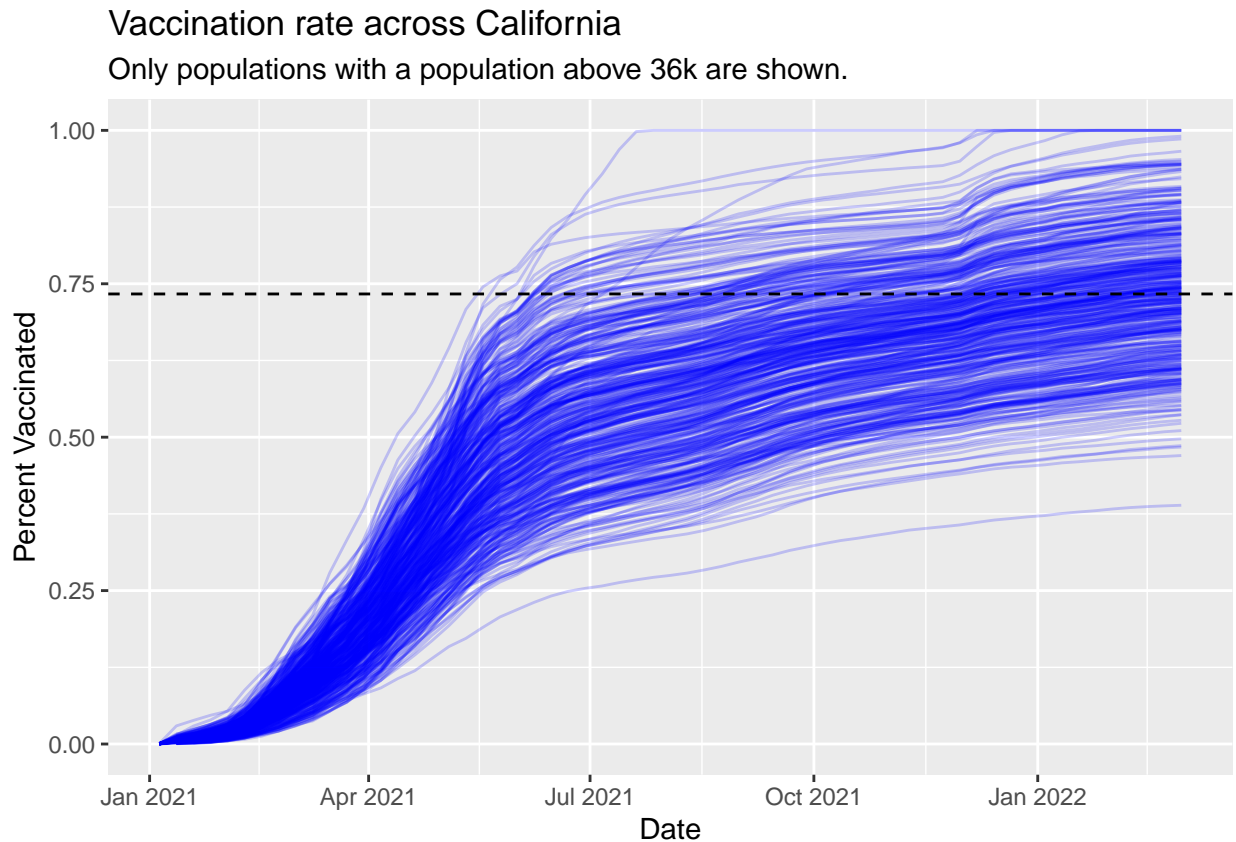
Answer: Calculated average value for ZIP code areas with a population as large as 92037 (La Jolla) as of date "2022-02-22" is 0.7334. Both 92040 and 92109 ZIP code areas are below the calculated average value (0.5513 and 0.7230 respectively).

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, 1) +
  labs(x="Date", y="Percent Vaccinated",
       title="Vaccination rate across California",
       subtitle="Only populations with a population above 36k are shown.") +
  geom_hline(yintercept = mean.36, linetype="dashed")
```

Answer:

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



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

Answer: I don't mind traveling for Spring break since the vaccination rate across CA has been above 75 percentile. However, I still prefer having remote/hybrid class.