Class17 Vaccination Mini Project

Monica Lin (PID: A15524235)

11/25/2021

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

Move the downloaded CSV file to the Class17 project directory, then read/import into an R object named vax. 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                    92395            San Bernardino San
Bernardino
## 2 2021-01-05                    93206                      Kern
Kern
## 3 2021-01-05                    91006                Los Angeles    Los
Angeles
## 4 2021-01-05                    91901                  San Diego      San
Diego
## 5 2021-01-05                    92230                  Riverside
Riverside
## 6 2021-01-05                    92662                     Orange
Orange
##   vaccine_equity_metric_quartile                 vem_source
## 1                              1 Healthy Places Index Score
## 2                              1 Healthy Places Index Score
## 3                              3 Healthy Places Index Score
## 4                              3 Healthy Places Index Score
## 5                              1 Healthy Places Index Score
## 6                              4 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1               35915.3                40888                       NA
## 2                1237.5                 1521                       NA
## 3               28742.7                31347                       19
```

```
## 4               15549.8                16905                        12
## 5                2320.2                 2526                        NA
## 6                2349.5                 2397                        NA
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                           NA                                     NA
## 2                           NA                                     NA
## 3                          873                                0.000606
## 4                          271                                0.000710
## 5                           NA                                     NA
## 6                           NA                                     NA
##   percent_of_population_partially_vaccinated
## 1                                         NA
## 2                                         NA
## 3                                   0.027850
## 4                                   0.016031
## 5                                         NA
## 6                                         NA
##   percent_of_population_with_1_plus_dose
## 1                                     NA
## 2                                     NA
## 3                               0.028456
## 4                               0.016741
## 5                                     NA
## 6                                     NA
##                                                                redacted
## 1 Information redacted in accordance with CA state privacy requirements
## 2 Information redacted in accordance with CA state privacy requirements
## 3                                                                    No
## 4                                                                    No
## 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?

The column "persons_fully_vaccinated" details the total number of people fully vaccinated.

Q2. What column details the Zip code tabulation area?

"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"
```

The earliest date in the dataset is 2021-01-05, by Year-month-date.

Q4. What is the latest date in this dataset?

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

```
## [1] "2021-11-23" "2021-11-23" "2021-11-23" "2021-11-23" "2021-11-23"
## [6] "2021-11-23"
```

The latest date in this dataset is 2021-11-23.

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

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

Data summary

Name	vax
Number of rows	82908
Number of columns	14

_____

Column type frequency:	
character	5
numeric	9

_____

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	47	0
local_health_jurisdiction	0	1	0	15	235	62	0
county	0	1	0	15	235	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	7.39	90001	922 57.75	936 58.50	953 80.50	976 35.0	▂▇▅▇▅▇▁
vaccine_equity_metric_quartile	4089	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	▇▇▁▅▇

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
age12_plus_population	0	1.00	18895.04	18993.94	0	1346.95	13685.10	31756.12	88556.7	
age5_plus_population	0	1.00	20875.24	21106.04	0	1460.50	15364.00	34877.00	101902.0	
persons_fully_vaccinated	8355	0.90	9585.35	11609.12	11	516.00	4210.00	16095.00	71219.0	
persons_partially_vaccinated	8355	0.90	1894.87	2105.55	11	198.00	1269.00	2880.00	20159.0	
percent_of_population_fully_vaccinated	8355	0.90	0.43	0.27	0	0.20	0.44	0.63	1.0	
percent_of_population_partially_vaccinated	8355	0.90	0.10	0.10	0	0.06	0.07	0.11	1.0	
percent_of_population_with_1_plus_dose	8355	0.90	0.51	0.26	0	0.31	0.53	0.71	1.0	

Q5. How many numeric columns are in this dataset?

9

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

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

## [1] 8355
```

There are 8355 NA values in that column.

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

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

## [1] 0.1007744
```

10.08% of `persons_fully_vaccinated` values are missing.

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

Optional.

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 makes life a lot easier when dealing with dates and times. Here is a quick example to get you started:

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

## Warning: package 'lubridate' was built under R version 4.1.2

##
## Attaching package: 'lubridate'

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

What is today's date?

```
today()

## [1] "2021-11-27"
```

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 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 mnay days have passed since the first vaccination reported in this dataset?

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

## Time difference of 326 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 322 days
```

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

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

## Time difference of 4 days
```

It has been 4 days since the last entry.

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

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

## [1] 47
```

There are 47 unique dates in the 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 thn load up this package to find the centroid of the La Jolla 92037 (i.e. UC San Diego) ZIP code area.

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

## Warning: package 'zipcodeR' was built under R version 4.1.2

# Find 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:

```
table(vax$county)

##
##                          Alameda          Alpine          Amador
Butte
##              235            2303              47             564
846
##        Calaveras          Colusa    Contra Costa       Del Norte       El
Dorado
##              846             329            2021             188
1034
##           Fresno           Glenn        Humboldt        Imperial
Inyo
##             2585             282            1645             705
470
##             Kern           Kings            Lake          Lassen     Los
```

```
Angeles
##            2303             329             658             611
13630
##          Madera           Marin        Mariposa       Mendocino
Merced
##             564            1316             376            1222
893
##           Modoc            Mono        Monterey            Napa
Nevada
##             517             329            1316             470
564
##          Orange          Placer          Plumas       Riverside
Sacramento
##            4136            1363             752            3290
2538
##      San Benito  San Bernardino       San Diego   San Francisco     San
Joaquin
##             188            4183            5029            1269
1504
## San Luis Obispo       San Mateo   Santa Barbara     Santa Clara      Santa
Cruz
##            1034            1363            1081            2726
799
##          Shasta          Sierra        Siskiyou          Solano
Sonoma
##            1222             329             987             705
1692
##      Stanislaus          Sutter          Tehama         Trinity
Tulare
##            1128             423             611             611
1551
##        Tuolumne         Ventura            Yolo            Yuba
##             611            1269             799             517
```

```
inds <- vax$county=="San Diego"
head(vax[inds, ])
```

```
##    as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 4  2021-01-05                    91901                 San Diego San Diego
## 14 2021-01-05                    91902                 San Diego San Diego
## 21 2021-01-05                    92011                 San Diego San Diego
## 22 2021-01-05                    92055                 San Diego San Diego
## 25 2021-01-05                    92067                 San Diego San Diego
## 33 2021-01-05                    92081                 San Diego San Diego
##    vaccine_equity_metric_quartile                 vem_source
## 4                               3 Healthy Places Index Score
## 14                              4 Healthy Places Index Score
## 21                              4 Healthy Places Index Score
## 22                              3    CDPH-Derived ZCTA Score
## 25                              4 Healthy Places Index Score
```

```
## 33                                2 Healthy Places Index Score
##    age12_plus_population age5_plus_population persons_fully_vaccinated
## 4                15549.8                16905                       12
## 14               16620.7                18026                       22
## 21               20503.6                23247                       NA
## 22               11548.0                11654                       NA
## 25                6973.9                 7480                       11
## 33               25558.0                27632                       14
##    persons_partially_vaccinated percent_of_population_fully_vaccinated
## 4                           271                                0.000710
## 14                          374                                0.001220
## 21                           NA                                      NA
## 22                           NA                                      NA
## 25                          241                                0.001471
## 33                          346                                0.000507
##    percent_of_population_partially_vaccinated
## 4                                    0.016031
## 14                                   0.020748
## 21                                         NA
## 22                                         NA
## 25                                   0.032219
## 33                                   0.012522
##    percent_of_population_with_1_plus_dose
## 4                                0.016741
## 14                               0.021968
## 21                                     NA
## 22                                     NA
## 25                               0.033690
## 33                               0.013029
##                                                                   redacted
## 4                                                                       No
## 14                                                                      No
## 21 Information redacted in accordance with CA state privacy requirements
## 22 Information redacted in accordance with CA state privacy requirements
## 25                                                                      No
## 33                                                                      No
```

Using the **dplyr** package and its **filter()** function, the code would look like this:

```
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] 5029
```

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
```

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?

```
which.max(sd$age12_plus_population)

## [1] 60

sd$zip_code_tabulation_area[23]

## [1] 92057
```

The San Diego County ZIP code area of 92057 has the largest 12+ population in this dataset.

Using **dplyr**, select all San Diego *"county"* entries on *"as_of_date"* "2021-11-09" and use this for the following questions.

```
sd.11.09 <- filter(vax, county=="San Diego" & as_of_date=="2021-11-09")
```

Q13. What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2021-11-09"?

```
mean(sd.11.09$percent_of_population_fully_vaccinated, na.rm=TRUE)

## [1] 0.6734714
```

The overall average "Percent of Population Fully Vaccinated" value is 67.34714%.

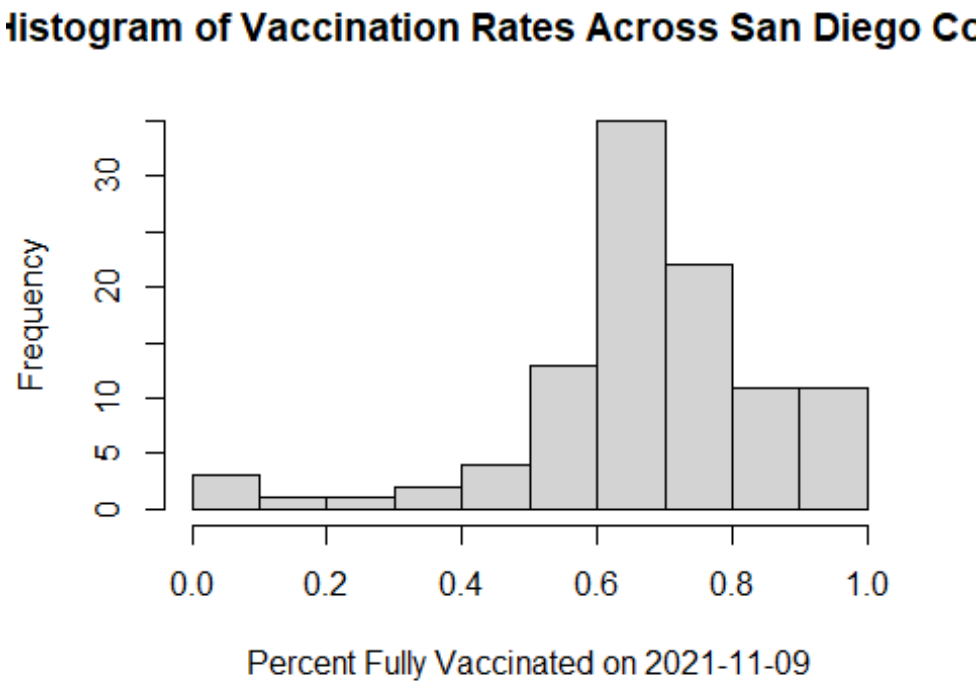
We can look at the 6-number summary.

```
summary(sd.11.09$percent_of_population_fully_vaccinated)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.01017 0.60805 0.67711 0.67347 0.76257 1.00000       4
```

Q14. Using either ggplot or base R graphics, make a summary figure that show the distribution of Percent of Population Fully Vaccinated values as of "2021-11-09"?

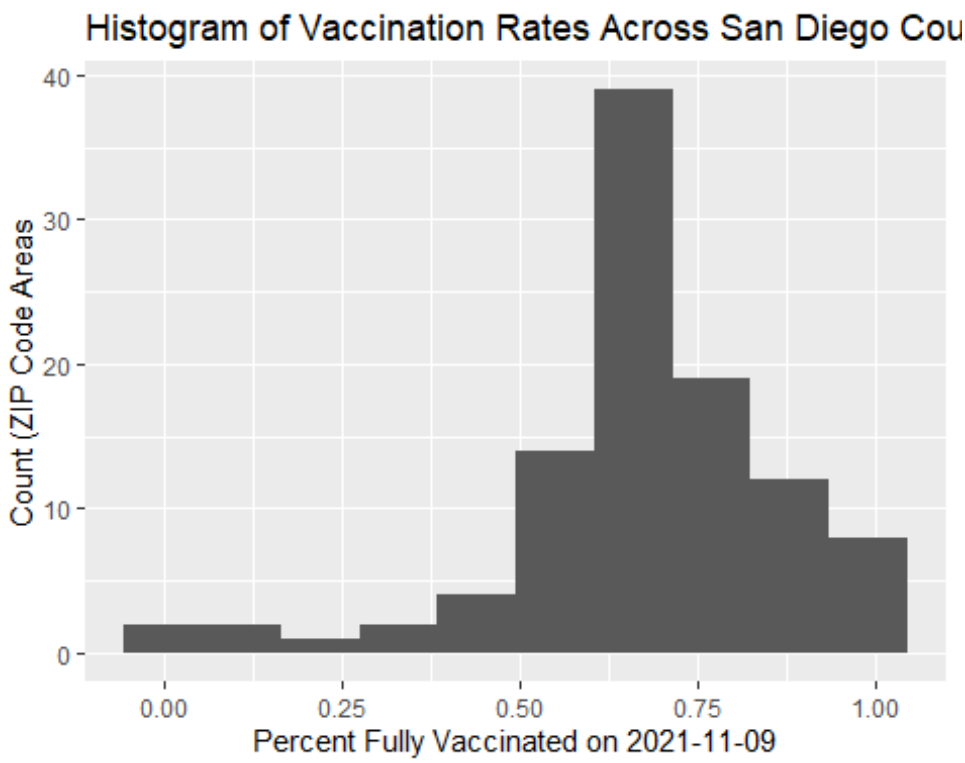
Using base R plots:

```
hist(sd.11.09$percent_of_population_fully_vaccinated,
     main="Histogram of Vaccination Rates Across San Diego County",
     xlab="Percent Fully Vaccinated on 2021-11-09",
     ylab="Frequency")
```



Using ggplot:

```
library(ggplot2)

ggplot(sd.11.09) +
  aes(percent_of_population_fully_vaccinated) +
  geom_histogram(bins=10) +
  labs(x="Percent Fully Vaccinated on 2021-11-09", y="Count (ZIP Code Areas",
       title="Histogram of Vaccination Rates Across San Diego County")

## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



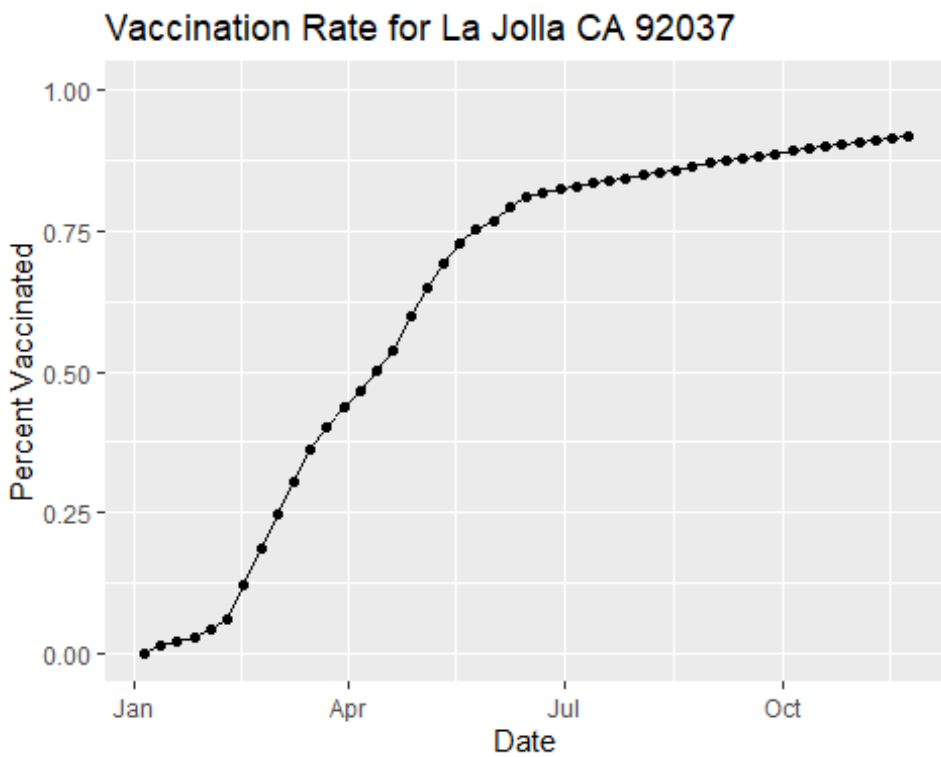
Focus on UCSD/La Jolla

UC San Diego resides in the 92037 ZIP code area nd 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",
       title="Vaccination Rate for La Jolla CA 92037")
```



This plot shows an initial slow roll out in January into Febuary (likely due to limited vaccine availability). This is followed with rapid ramp up until a clear slowing trend from June time onward. Interpertation beyond this requies 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? Etc.

Comparing 92037 to other 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* "2021-11-16".

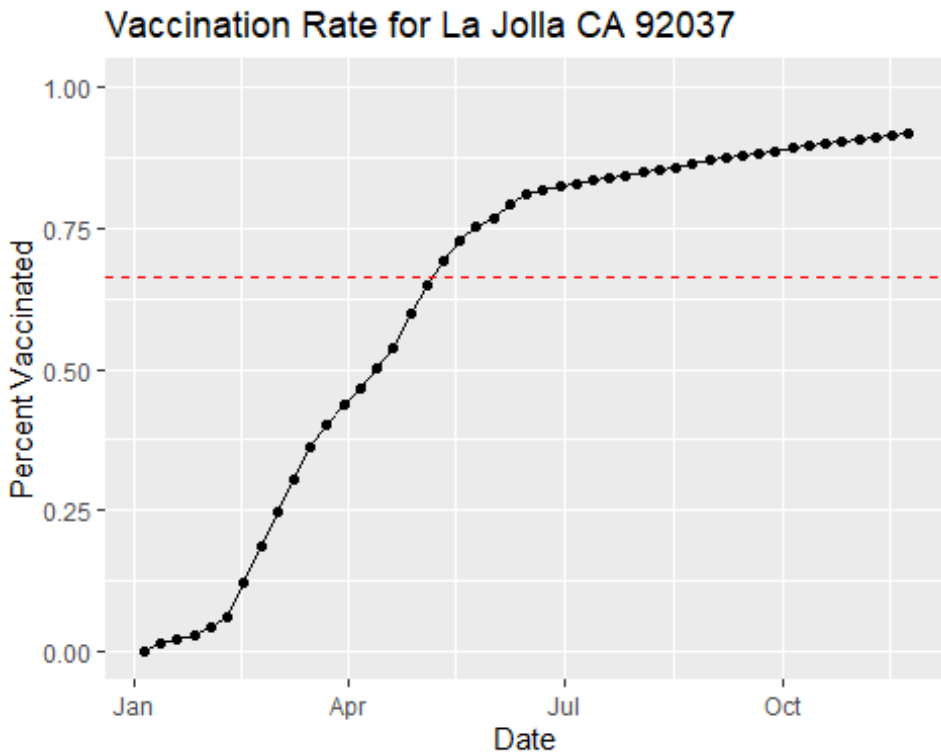
```
# Subset to all CA areas with a population as large as 92037
vax.36 <- filter(vax, age5_plus_population > 36144 &
                    as_of_date == "2021-11-16")

head(vax.36)

##   as_of_date zip_code_tabulation_area local_health_jurisdiction
county
## 1 2021-11-16                    92020                 San Diego      San
Diego
## 2 2021-11-16                    92563                 Riverside
Riverside
## 3 2021-11-16                    92806                    Orange
Orange
```

```
## 4 2021-11-16                    93291                    Tulare
Tulare
## 5 2021-11-16                    92335            San Bernardino San
Bernardino
## 6 2021-11-16                    92618                    Orange
Orange
##   vaccine_equity_metric_quartile                 vem_source
## 1                              2 Healthy Places Index Score
## 2                              3 Healthy Places Index Score
## 3                              2 Healthy Places Index Score
## 4                              1 Healthy Places Index Score
## 5                              1 Healthy Places Index Score
## 6                              4 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1               49284.5                54991                     35128
## 2               55897.8                63794                     36051
## 3               33050.9                36739                     24810
## 4               46879.7                54254                     27936
## 5               79670.3                91867                     49820
## 6               40348.0                44304                     39695
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                         5161                                0.638795
## 2                         4224                                0.565116
## 3                         2355                                0.675304
## 4                         4012                                0.514911
## 5                         5970                                0.542306
## 6                         3936                                0.895969
##   percent_of_population_partially_vaccinated
## 1                                   0.093852
## 2                                   0.066213
## 3                                   0.064101
## 4                                   0.073948
## 5                                   0.064985
## 6                                   0.088841
##   percent_of_population_with_1_plus_dose redacted
## 1                               0.732647       No
## 2                               0.631329       No
## 3                               0.739405       No
## 4                               0.588859       No
## 5                               0.607291       No
## 6                               0.984810       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* "2021-11-16". Add this as a straight horizontal line to your plot from above with the `geom_hline()` function.

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

```
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",
       title="Vaccination Rate for La Jolla CA 92037") +
  geom_hline(yintercept=vaccination.36, color="red", linetype="dashed")
```

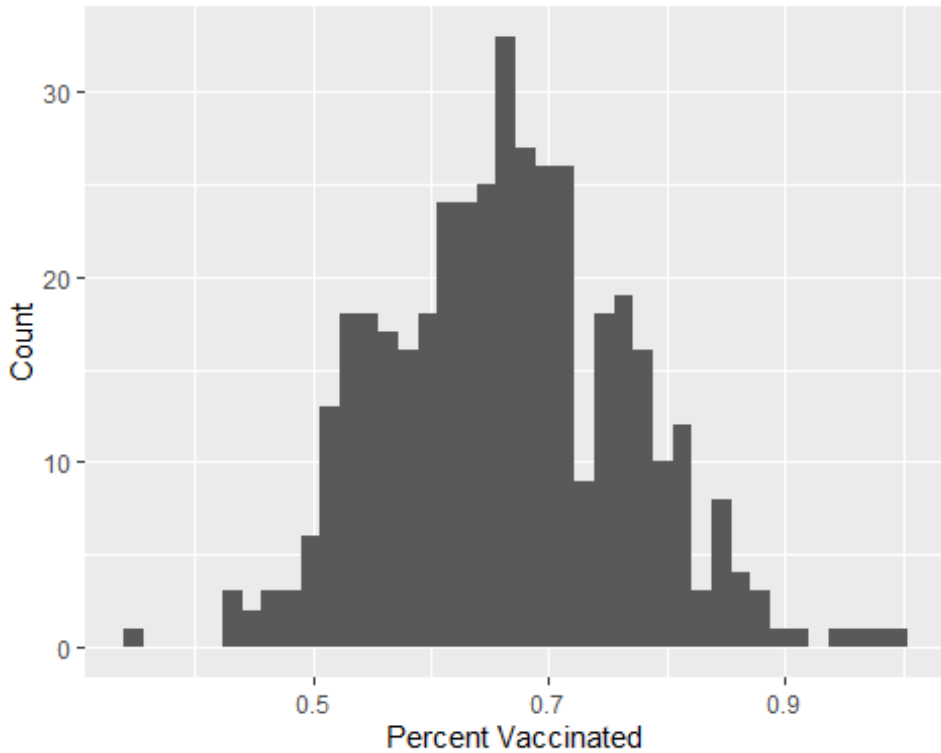


Q17. What is the 6 number summary (Min, 1st Qu., Median, Mean, 3rd Qu., and Max) of the *"Percent of Population Fully Vaccinated"* values for ZIP code areas with a population as large as 92037 (La Jolla) *as_of_date* "2021-11-16"?

```
summary(vax.36$percent_of_population_fully_vaccinated)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3529  0.5905  0.6662  0.6640  0.7298  1.0000
```

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

```
ggplot(vax.36) + aes(percent_of_population_fully_vaccinated) +
  geom_histogram(bins=40) + labs(x="Percent Vaccinated", y="Count")
```



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 == "2021-11-16") %>%
  filter(zip_code_tabulation_area=="92109") %>%
  select(percent_of_population_fully_vaccinated)

##   percent_of_population_fully_vaccinated
## 1                                0.68863

vax %>% filter(as_of_date == "2021-11-16") %>%
  filter(zip_code_tabulation_area=="92040") %>%
  select(percent_of_population_fully_vaccinated)

##   percent_of_population_fully_vaccinated
## 1                               0.521047
```

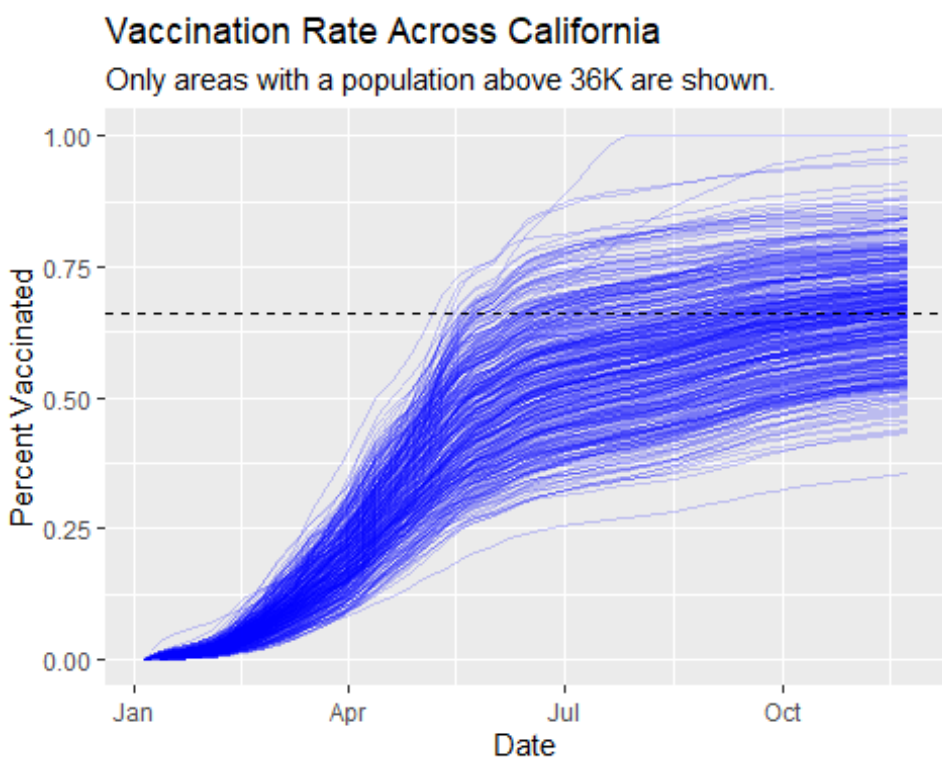
The 92109 ZIP code area is above the average value of 0.6630. However, the 92040 ZIP code area is below the average value.

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 a population above 36K are shown.") +
  geom_hline(yintercept=0.66, linetype="dashed")
```

```
## Warning: Removed 176 row(s) containing missing values (geom_path).
```



Q21. How do you feel about traveling for Thanksgiving and meeting for in-person class next Week?

With the detection of the omicron variant, which is more transmittable than the delta variant, and the combination of the lower-than-expected vaccination rates uncovered in this activity, I feel hesitant about meeting for in-person class next week. Traveling by car is safe enough, but traveling by plane for Thanksgiving is slightly concerning to me.