Vaccination Rate Mini Project

Matt Hashimoto

12/3/2021

Getting Started

Let's first start by loading our data from the .csv file:

```
# 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                    92091                 San Diego    San
Diego
## 2 2021-01-05                    92116                 San Diego    San
Diego
## 3 2021-01-05                    95360                Stanislaus
Stanislaus
## 4 2021-01-05                    94564              Contra Costa Contra
Costa
## 5 2021-01-05                    95501                  Humboldt
Humboldt
## 6 2021-01-05                    95492                    Sonoma
Sonoma
##   vaccine_equity_metric_quartile                 vem_source
## 1                              4    CDPH-Derived ZCTA Score
## 2                              3 Healthy Places Index Score
## 3                              1 Healthy Places Index Score
## 4                              4 Healthy Places Index Score
## 5                              2 Healthy Places Index Score
## 6                              4 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1                1238.3                 1303                       NA
## 2               30255.7                31673                       45
## 3               10478.5                12301                       NA
## 4               17033.0                18381                       NA
## 5               20566.6                22061                       NA
## 6               25076.9                28024                       NA
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                           NA                                     NA
## 2                          898                                0.001421
## 3                           NA                                     NA
## 4                           NA                                     NA
## 5                           NA                                     NA
```

```
## 6                              NA                                        NA
##   percent_of_population_partially_vaccinated
## 1                                         NA
## 2                                   0.028352
## 3                                         NA
## 4                                         NA
## 5                                         NA
## 6                                         NA
##   percent_of_population_with_1_plus_dose
## 1                                     NA
## 2                                0.029773
## 3                                     NA
## 4                                     NA
## 5                                     NA
## 6                                     NA
##                                                                  redacted
## 1 Information redacted in accordance with CA state privacy requirements
## 2                                                                     No
## 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?"

Column 9, titled "persons_fully_vaccinated".

Q2

"What column details the Zip code tabulation area?"

Column 2, titled "zip_code_tabulation_area".

Q3

"What is the earliest date in this dataset?"

This can be found by looking at the first entry in the "as_of_date" column:

```
# View the first entry in the as_of_date column
vax$as_of_date[1]
```

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

Thus, the earliest date is January 5th, 2021.

Q4

"What is the latest date in this dataset?"

Similarly to the last question, this can be found by looking at the last entry in the "as_of_date" column:

```
vax$as_of_date[length(vax$as_of_date)]

## [1] "2021-11-30"
```

Thus, the latest date is November 30th, 2021.

Let's try calling the skim function to get a better idea of what's in the dataset:

```
# Call the skim function
skimr::skim(vax)
```

Data summary

Name	vax
Number of rows	84672
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	48	0
local_health_jurisdiction	0	1	0	15	240	62	0
county	0	1	0	15	240	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	181 7.39	90001	92257.5	93658.50	95380.50	97635.0	▂▆▇▂▁

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
vaccine_equity_metric_quartile	4176	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	▇▇▁▇▇
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	8472	0.90	9709.47	11714.06	11	526.00	4309.50	16316.00	71552.0	▇▂▁▁▁
persons_partially_vaccinated	8472	0.90	1891.41	2100.88	11	197.00	1268.50	2874.00	20158.0	▇▁▁▁▁
percent_of_population_fully_vaccinated	8472	0.90	0.43	0.27	0	0.21	0.45	0.63	1.0	▅▅▇▅▃
percent_of_population_partially_vaccinated	8472	0.90	0.10	0.10	0	0.06	0.07	0.11	1.0	▇▁▁▁▁
percent_of_population_with_1_plus_dose	8472	0.90	0.51	0.26	0	0.31	0.54	0.71	1.0	▃▃▅▃▅

Q5

"How many numeric columns are in this dataset?"

As seen from the skim results, there are 9 numeric columns.

Q6

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

The "n_missing" column shows that there are 8472 NA values in the "persons_fully_vaccinated" column.

Q7

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

```
# 8472 missing values out of 84672
8472 / 84672
```

```
## [1] 0.1000567
```

10% of the values are missing.

Q8

"[Optional]: Why might this data be missing?"

This data may be missing because there is no method of collecting data from specific zip codes. As mentioned earlier in the lab document, certain institutions or organizations may have no obligation or reason to report their vaccination data, and certain zip codes may be entirely managed by these institutions or organizations.

Working With Dates

Let's use the lubridate library to help us deal with dates:

```
library(lubridate)

##
## Attaching package: 'lubridate'

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

Check today's date:

```
today()

## [1] "2021-12-03"
```

Let's convert our dates into a lubridate format to make analysis easier:

```
# Speciffy that we are using the Year-mont-day format
vax$as_of_date <- ymd(vax$as_of_date)
```

Now we can use lubridate functions to check things like how many days have passed since the first data was collected:

```
# Check time since first measurement
today() - vax$as_of_date[1]

## Time difference of 332 days
```

We can also calculate how much time the data spans:

```
# Check time span
vax$as_of_date[nrow(vax)] - vax$as_of_date[1]

## Time difference of 329 days
```

Q9

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

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

## Time difference of 3 days
```

3 days have passed since the last update.

Q10

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

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

## [1] 48
```

There are 48 unique dates in the dataset.

Working With ZIP Codes

Let's load in the zipcodeR library:

```
# Load the zipcodeR library
library(zipcodeR)
```

Next let's find the centroid of the 92037 zip code area (UCSD):

```
# Find centroid of the 92037 zip code
geocode_zip('92037')

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

We can also calculate the distance between any two zip codes in miles:

```
# Distance in miles
zip_distance('92037','92109')

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

We can also pull census data about zip codes:

```
# Pull census data
reverse_zipcode(c('92037', "92109") )

## # A tibble: 2 × 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 to pull census data for all the zip codes we may 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

We can restrict ourselves to San Diego county using base R:

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

Or we could use the dplyr library:

```
# Load library
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

# Filter just results from SD
sd <- filter(vax, county == "San Diego")
nrow(sd)

## [1] 5136
```

The dplyr package is more convenient when trying to subset across multiple criteria:

```
# All SD counties with populations over 10000
sd.10 <- filter(vax, county == "San Diego" &
                age5_plus_population > 10000)
```

Q11

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

```
# Check for uniqueness
length(unique(sd$zip_code_tabulation_area))

## [1] 107
```

107 distinct zip codes are listed for SD county.

Q12

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

```
# Check for max population value
sd$zip_code_tabulation_area[which.max(sd$age12_plus_population)]

## [1] 92154
```

The 92154 area has the largest 12+ population.

```
# All data for Nov 16
sd.nov16 <- filter(vax, county == "San Diego" &
                 as_of_date == "2021-11-16")
```

Q13

"What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2021-11-16"?"

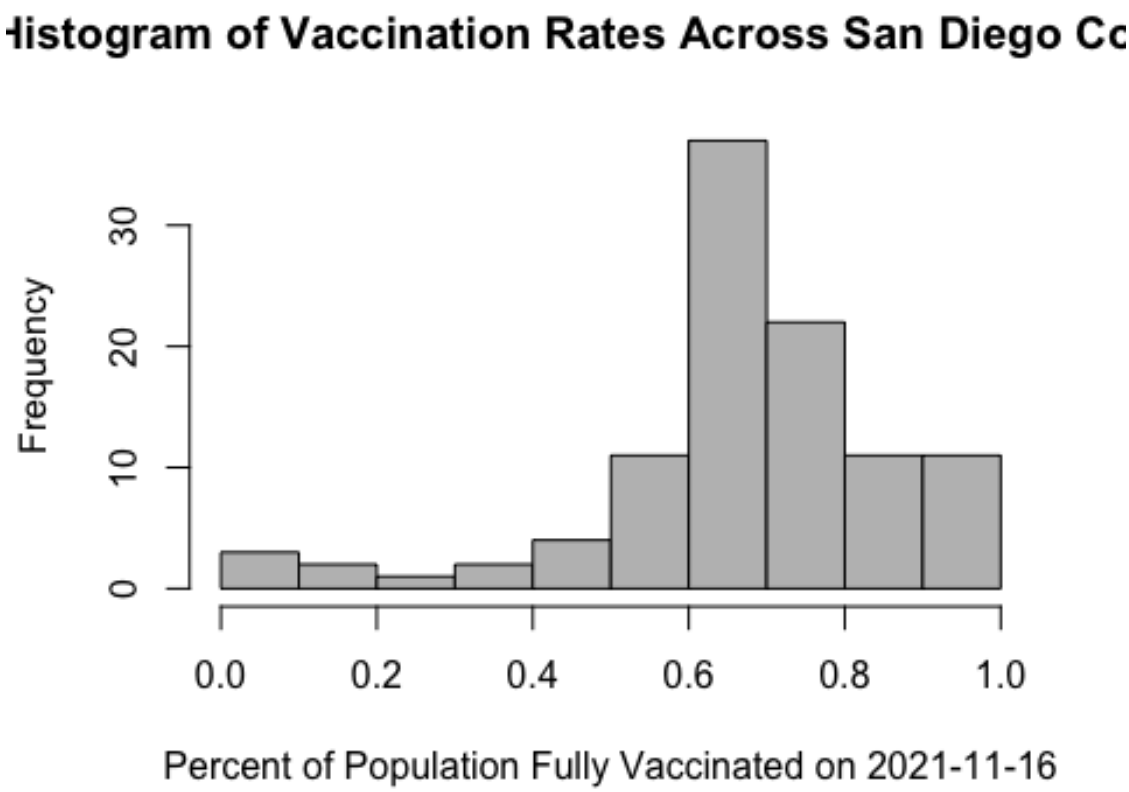
```
# Average percent of population fully vaccinated
mean(sd.nov16$percent_of_population_fully_vaccinated, na.rm = TRUE)

## [1] 0.6722183
```

The average percent of population fully vaccinated is 67.22%.

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 "2021-11-16"?"

```
# Plot distribution of percent fully vaccinated
hist(sd.nov16$percent_of_population_fully_vaccinated,
     main = "Histogram of Vaccination Rates Across San Diego County",
     xlab = "Percent of Population Fully Vaccinated on 2021-11-16",
     col = "gray")
```



Focus on UCSD/La Jolla

Let's filter to the UCSD area zip code:

```
# Filter to UCSD zip code and check 5+ population
ucsd <- filter(sd, zip_code_tabulation_area == "92037")
ucsd[1,]$age5_plus_population

## [1] 36144
```

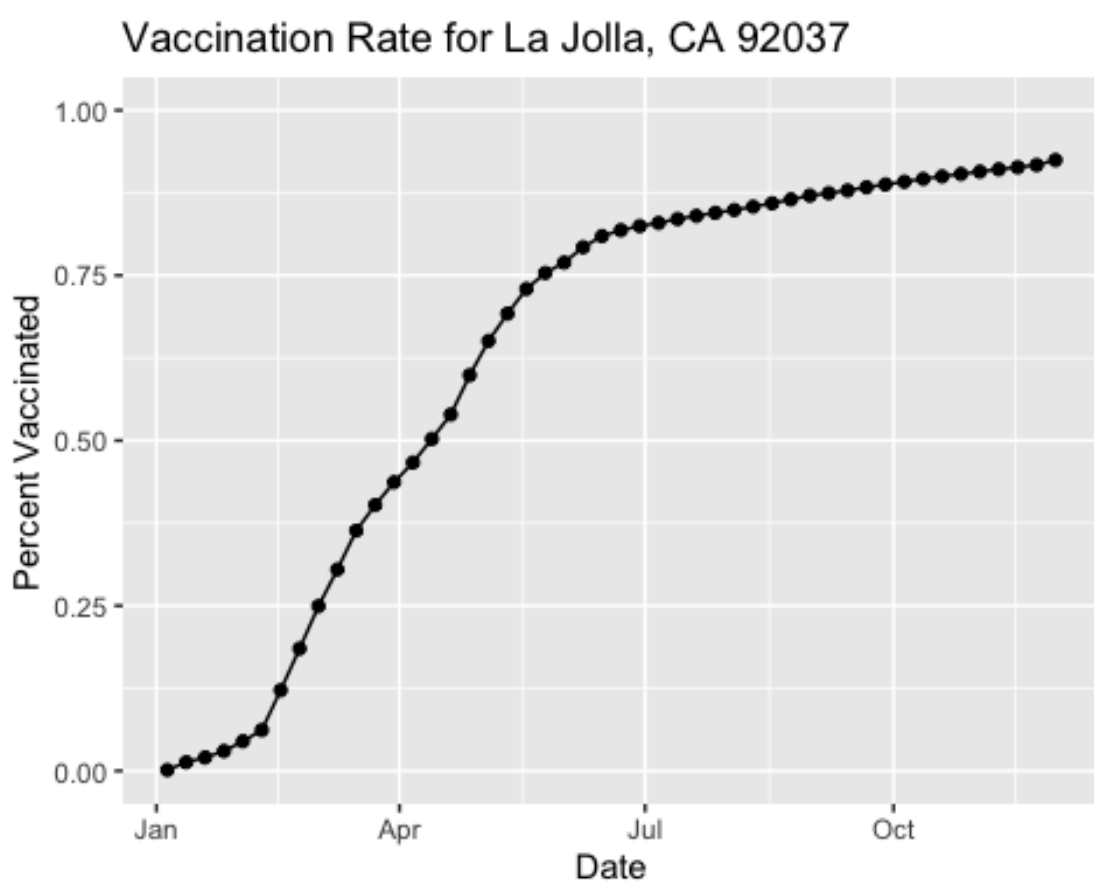
Q15

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

```
# Load ggplot library
library(ggplot2)

# Use ggplot to create a graph
ggplot(ucsd) +
  aes(ucsd$as_of_date,
      ucsd$percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group = 1) +
```

```
  ylim(c(0,1)) +
  labs(x = "Date", y = "Percent Vaccinated") +
  ggtitle("Vaccination Rate for La Jolla, CA 92037")
```

```
## 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.
```



Comparing 92037 to Other Similarly Sized Areas

Let's filter our vaccination data once again to data at least as large as the population in 92037:

```r
# 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)
```

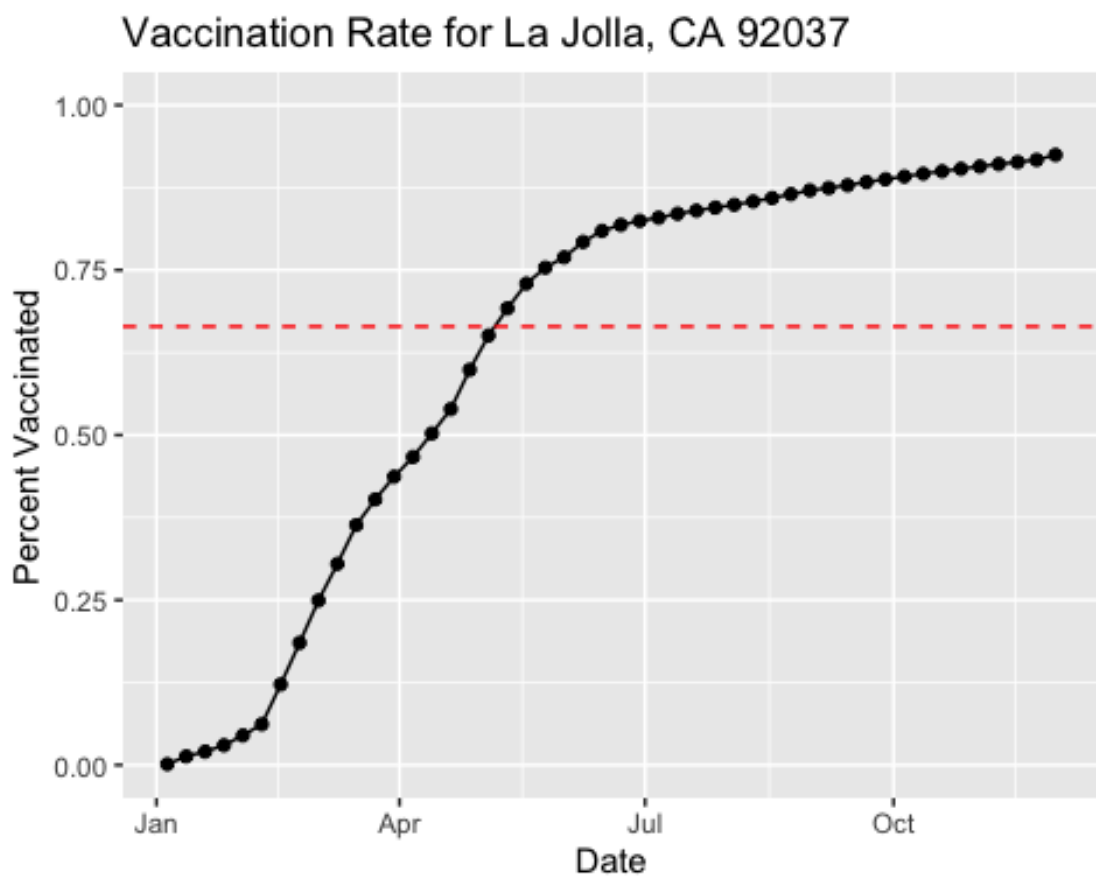
```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction
county
## 1 2021-11-16                    92345            San Bernardino San
Bernardino
## 2 2021-11-16                    92553                 Riverside
Riverside
## 3 2021-11-16                    92058                 San Diego      San
Diego
## 4 2021-11-16                    91786            San Bernardino San
Bernardino
## 5 2021-11-16                    92507                 Riverside
Riverside
## 6 2021-11-16                    93021                   Ventura
Ventura
##   vaccine_equity_metric_quartile                 vem_source
## 1                              1 Healthy Places Index Score
## 2                              1 Healthy Places Index Score
## 3                              1 Healthy Places Index Score
## 4                              2 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               66047.5                75539                     35432
## 2               61770.8                70472                     37411
## 3               34956.0                39695                     14023
## 4               45602.3                50410                     30834
## 5               51432.5                55253                     31939
## 6               32753.7                36197                     24918
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                         4389                                0.469056
## 2                         4846                                0.530863
## 3                         2589                                0.353269
## 4                         3132                                0.611664
## 5                         3427                                0.578050
## 6                         2012                                0.688400
##   percent_of_population_partially_vaccinated
## 1                                   0.058102
## 2                                   0.068765
## 3                                   0.065222
## 4                                   0.062131
## 5                                   0.062024
## 6                                   0.055585
##   percent_of_population_with_1_plus_dose redacted
## 1                               0.527158       No
```

```
## 2                                  0.599628       No
## 3                                  0.418491       No
## 4                                  0.673795       No
## 5                                  0.640074       No
## 6                                  0.743985       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?"

```
# Calculate mean
mean(vax.36$percent_of_population_fully_vaccinated, na.rm = TRUE)

## [1] 0.6645132

# Add line to plot
ggplot(ucsd) +
  aes(ucsd$as_of_date,
      ucsd$percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group = 1) +
  ylim(c(0,1)) +
  labs(x = "Date", y = "Percent Vaccinated") +
  ggtitle("Vaccination Rate for La Jolla, CA 92037") +
  geom_hline(yintercept = mean(vax.36$percent_of_population_fully_vaccinated,
                       na.rm = TRUE),
             color = "red", linetype = "dashed")

## 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.
```



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"?"

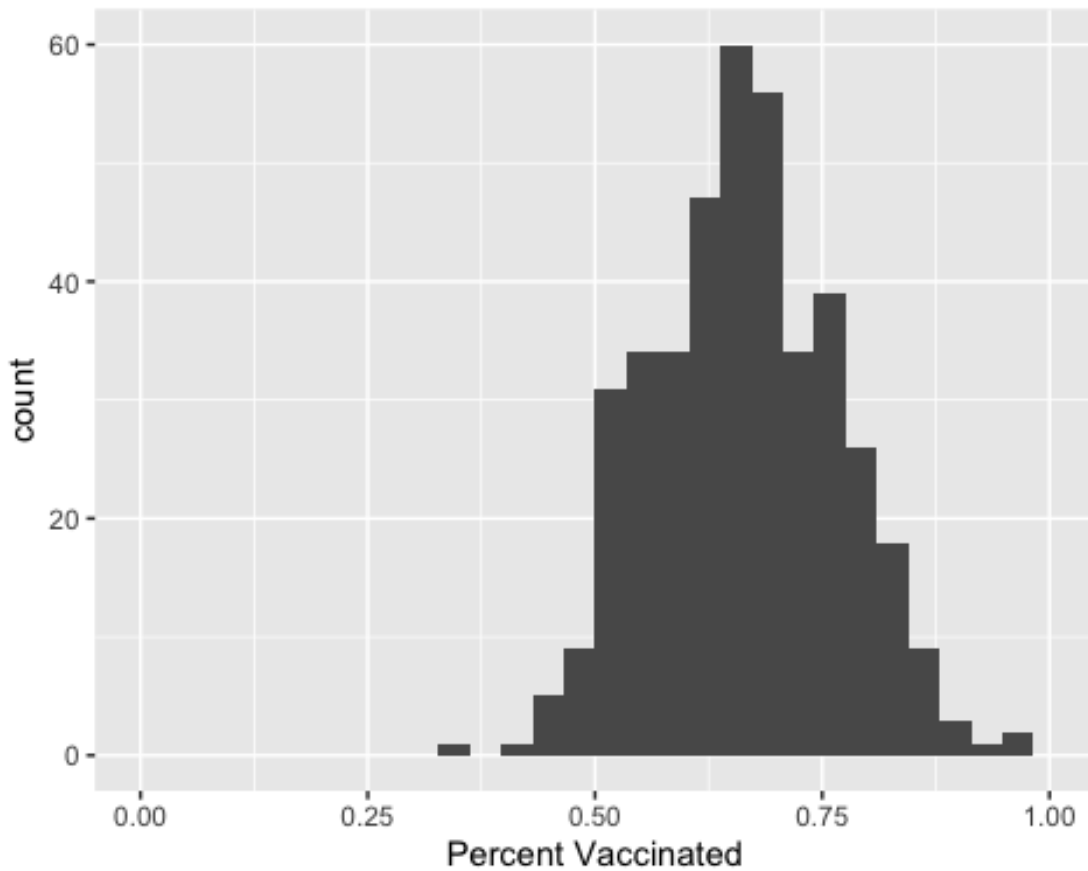
```
# Use fivenum to get min, 1st qu, median, 3rd qu, and max
fivenum(vax.36$percent_of_population_fully_vaccinated)

## [1] 0.353269 0.591029 0.666919 0.731112 1.000000

# Use mean()
mean(vax.36$percent_of_population_fully_vaccinated)

## [1] 0.6645132
```

Q18

"Using ggplot generate a histogram of this data."

```
ggplot(vax.36) +
  aes(vax.36$percent_of_population_fully_vaccinated) +
  geom_histogram() +
```

```
  xlim(0,1) +
  labs(x = "Percent Vaccinated")

## Warning: Use of `vax.36$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 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?"

```
# The average value
mean(vax.36$percent_of_population_fully_vaccinated)

## [1] 0.6645132

# Check 92109
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                                0.68912
```

```r
# Check 92040
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                                0.52142
```

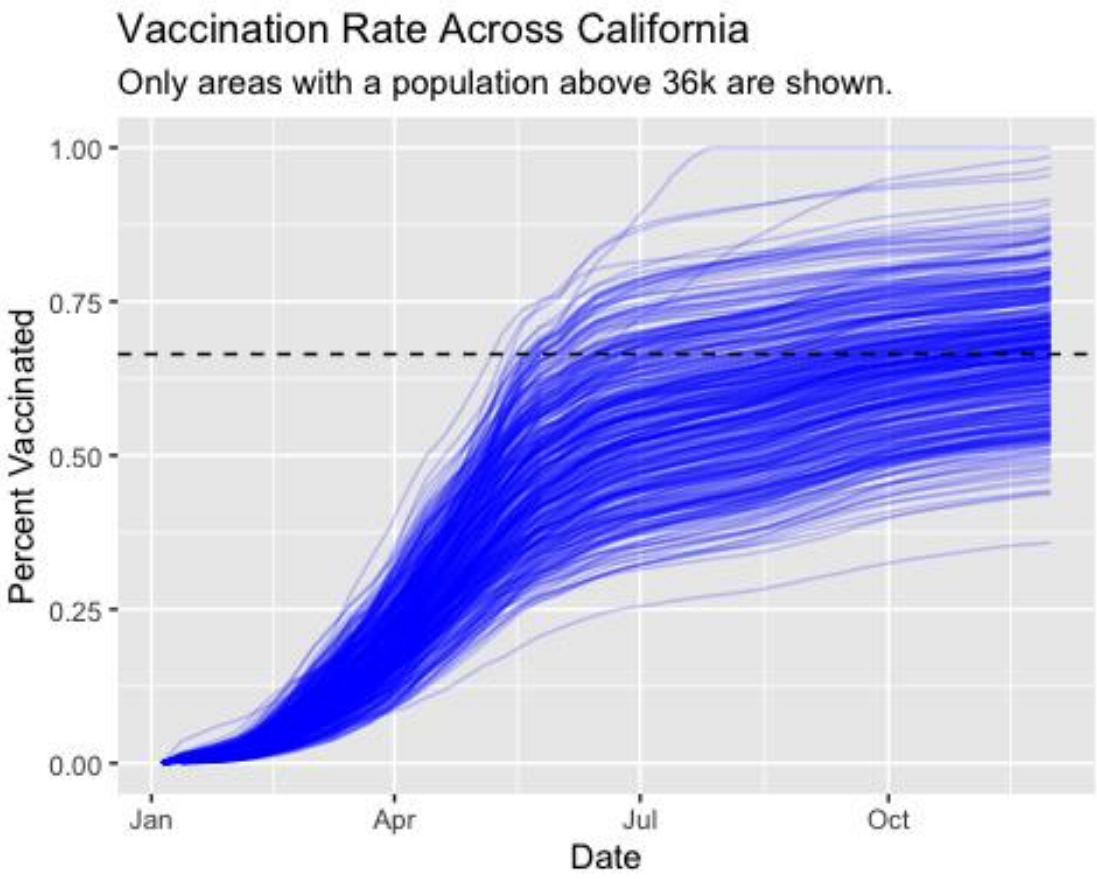
As you can see, the 92109 zip code is above the average vaccination percentage, while the 92040 zip code is below.

Q20

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

```r
# Filter data for all days
vax.36.all <- filter(vax, age5_plus_population > 36144)

# Plot with ggplot
ggplot(vax.36.all) +
  aes(vax.36.all$as_of_date,
      vax.36.all$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 areas with a population above 36k are shown.") +
  geom_hline(yintercept = mean(vax.36$percent_of_population_fully_vaccinated,
                               na.rm = TRUE),
             linetype = "dashed")
```

```
## Warning: Use of `vax.36.all$as_of_date` is discouraged. Use `as_of_date`
## instead.

## Warning: Use of `vax.36.all$percent_of_population_fully_vaccinated` is
## discouraged. Use `percent_of_population_fully_vaccinated` instead.

## Warning: Removed 177 row(s) containing missing values (geom_path).
```


Vaccination Rate Across California
Only areas with a population above 36k are shown.
1.00
0.75
0.50
0.25
0.00
Percent Vaccinated
Jan
Apr
Jul
Oct
Date