vaccine mini project

Soobin (PID:A15201229)

3/3/2022

```r
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?

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

> Q2. What column details the Zip code tabulation area?

"zip_code_tabulation_area" details the Zip code tabulation area.

> Q3. What is the earliest date in this dataset?

```
vax$as_of_date[1]
```

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

The earliest date in this dataset is `vax$as_of_date[1]`.

> Q4. What is the latest date in this dataset?

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

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

The latest date in this dataset is `vax$as_of_date[nrow(vax)]`.

Skim summarizes the data sets.

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

Table 1: Data summary

Name	vax
Number of rows	107604
Number of columns	15
Column type frequency:	
character	5
numeric	10

Table 1: Data summary

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_quartile	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	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	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_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	

> Q5. How many numeric columns are in this dataset?

There are 9 numeric columns because zipcode should not be used as a numeric value.

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

```
no.na <- sum( is.na( vax$persons_partially_vaccinated ) )
no.na
```

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

There are 18338 NA values in the persons_fully_vaccinated column.

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

```
round( no.na / nrow(vax), 2 )
```

```
## [1] 0.17
```

17% of persons_fully_vaccinated values are missing.

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

This data might be missing because people did not get their vaccines and reported to CDC.

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:

Lubridate works with dates (i.e. do math).

```
library(lubridate)
```

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

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

```
age <- today() - ymd("1998-04-21")
age
```

```
## Time difference of 8717 days
```

```
time_length(age, "year")
```

```
## [1] 23.86585
```

We cannot subtract vax$as_of_date[1] from today() because as_of_date is written in character function.

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

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

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

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

```
## Time difference of 2 days
```

2 days have passed since the last update of the 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
```

There are 61 unique dates in the dataset.

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.

dplyr package is used to work with data.

```
sd <- vax[vax$county == "San Diego", ]
dim(sd)
```

```
## [1] 6527   15
```

An often more convenient way to do this type of "filtering" (a.k.a. subsetting) is with the **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")
dim(sd)
```

```
## [1] 6527   15
```

> 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?

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

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

92154 San Diego County Zip code area has the largest 12+ Population in this dataset.

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 "2022-03-01"?

```
# Filter to the day
sd.latest <- filter(sd, as_of_date == "2022-03-01")
mean( sd.latest$percent_of_population_fully_vaccinated, na.rm = TRUE)
```

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

The overall average “Percent of Population Fully Vaccinated” value for all San Diego “County” as of “2022-03-01” is 70.53%.

> 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”?

```
summary(sd.latest$percent_of_population_fully_vaccinated)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
## 0.01017 0.65132 0.72452 0.70529 0.82567 1.00000       1
```

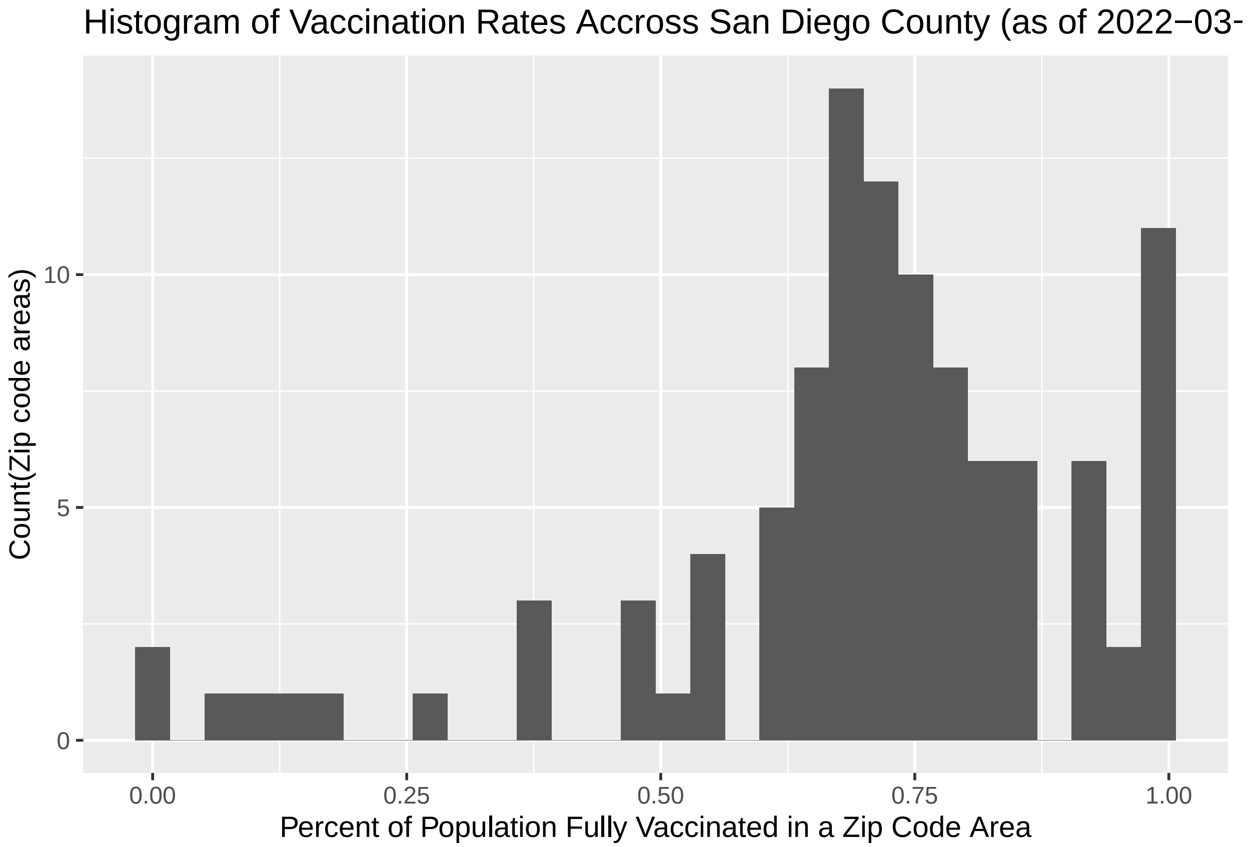
```
library(ggplot2)
```

```
ggplot(sd.latest) + aes(sd.latest$percent_of_population_fully_vaccinated) + geom_histogram() + labs(x =
```

```
## Warning: Use of `sd.latest$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 1 rows containing non-finite values (stat_bin).
```



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.

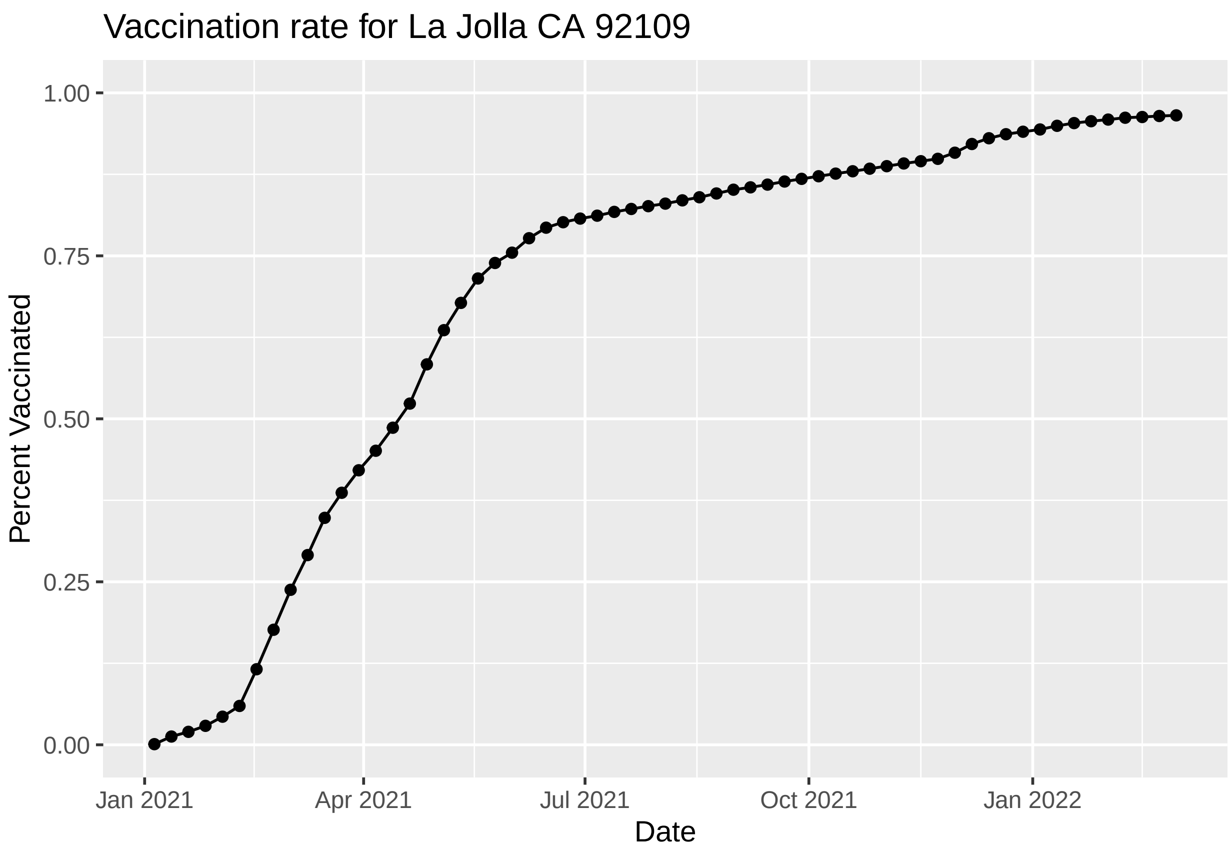
```
ucsd <- filter(sd, zip_code_tabulation_area == "92037")
ucsd$age5_plus_population[1]
```

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

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

```
baseplot <- ggplot(ucsd) + aes(as_of_date, percent_of_population_fully_vaccinated) + geom_point() + geor

baseplot
```



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 "2022-03-01".

> 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 "2022-03-01". Add this as a straight horizontal line to your plot from above with the geom_hline() function?

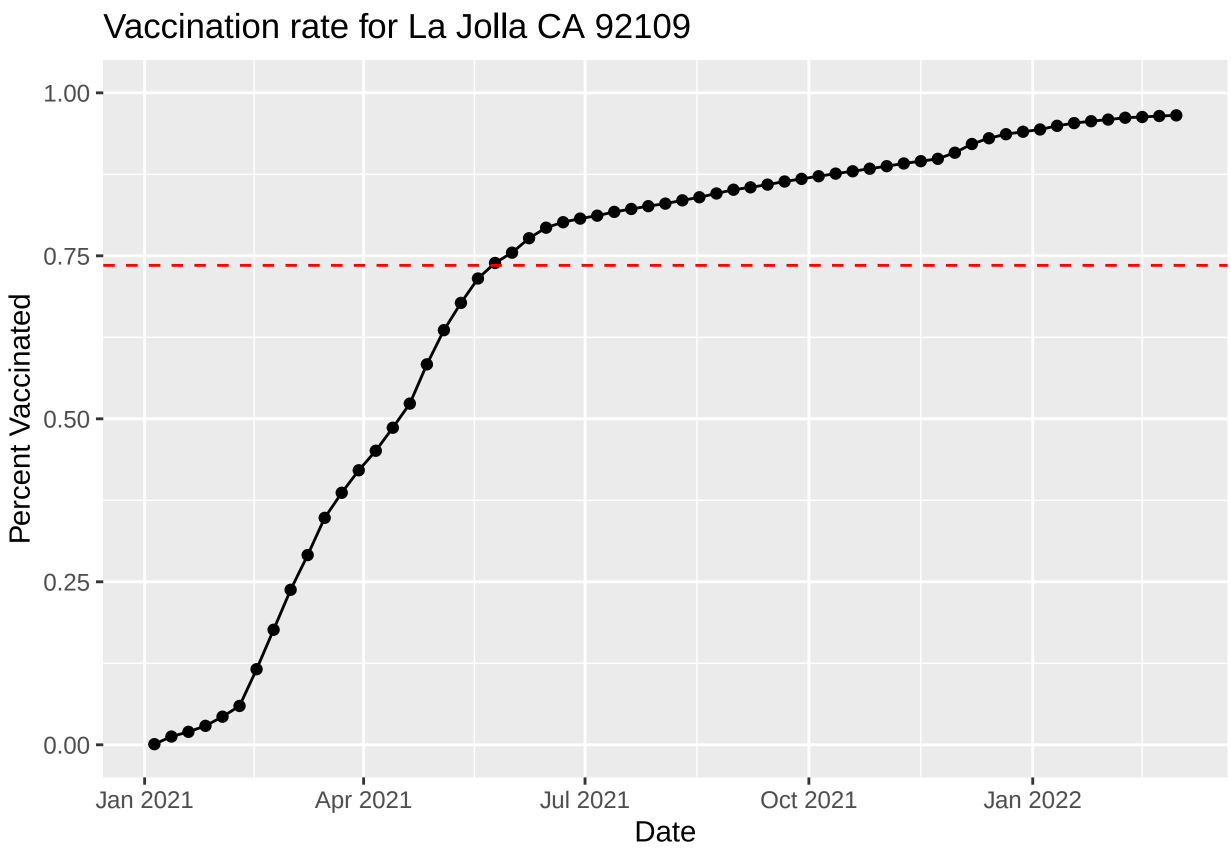
```
vax.36 <- filter(vax, age5_plus_population > 36144 & as_of_date == "2022-03-01")

vax.36.mean <- mean( vax.36$percent_of_population_fully_vaccinated, na.rm = T )

vax.36.mean
```

```
## [1] 0.7353974
```

```
baseplot + geom_hline( yintercept = vax.36.mean, linetype=2, col="red" )
```



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 "2022-03-01"?

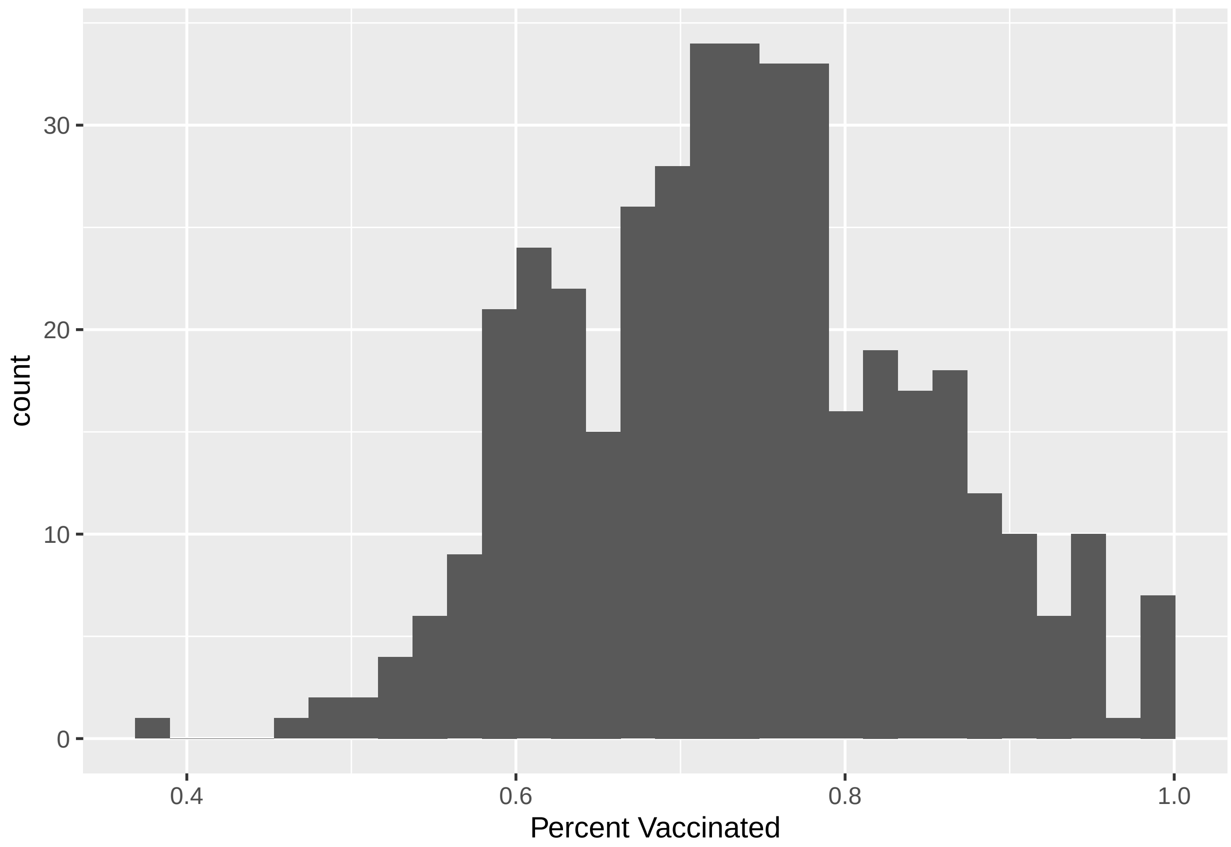
```
summary(vax.36$percent_of_population_fully_vaccinated)
```

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

Q18. Using ggplot generate a histogram of this data.

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

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Q19. Is the 92109 and 92040 ZIP code areas above or below the average value you calculated for all these above?

```
zip_92040 <- vax %>% filter(as_of_date == "2022-03-01") %>%
  filter(zip_code_tabulation_area=="92040") %>%
  select(percent_of_population_fully_vaccinated)

zip_92040
```

```
##   percent_of_population_fully_vaccinated
## 1                               0.551981
```

```
zip_92109 <- vax %>% filter(as_of_date == "2022-03-01") %>%
  filter(zip_code_tabulation_area=="92109") %>%
  select(percent_of_population_fully_vaccinated)

zip_92109
```

```
##   percent_of_population_fully_vaccinated
## 1                               0.723778
```

```
zip_92109 > vax.36.mean
```

```
##      percent_of_population_fully_vaccinated
## [1,]                                  FALSE
```

```
zip_92040 > vax.36.mean
```

```
##      percent_of_population_fully_vaccinated
## [1,]                                   FALSE
```

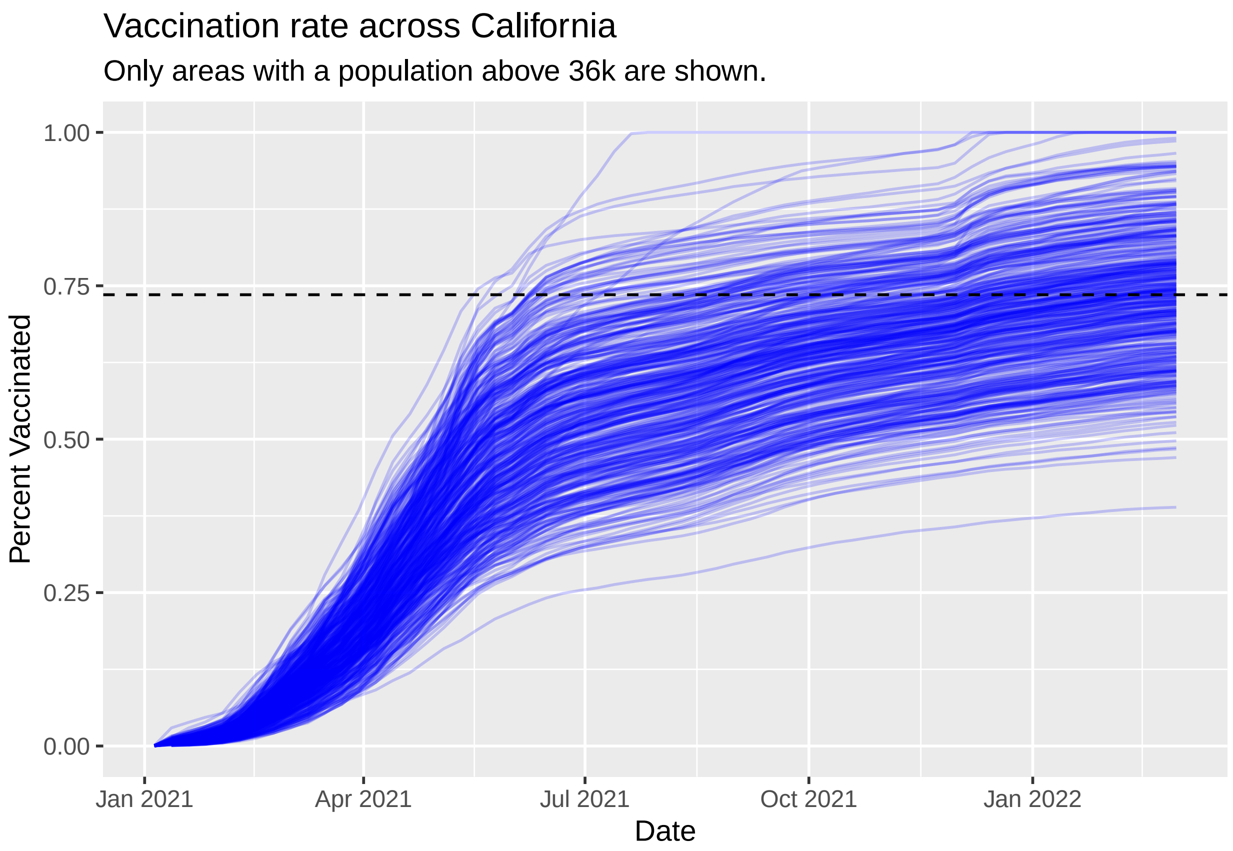
Both the 92109 and 92040 ZIP code areas are below the average value I calculated for all these above.

> 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(x=as_of_date, y=percent_of_population_fully_vaccinated, group=zip_code_tabulati
```

```
## 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?

A lot of areas are not fully vaccinated than I expected. But I still feel pretty safe traveling around because so many people got omicron during winter that I feel like most people would be either vaccinated or have immunity now. Of course, I am concerned but I am looking foward to travel a bit and join in-person classes.