COVID-19 Vaccination Rates

Kiley Hooker (PID: A15441609)

3/3/2022

Read our input data

Here we downloaded 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

```
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
```

```
#ymd(vax$as_of_date)
```

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? 2021-01-05

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

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

Q4. What is the latest date in this dataset? 2022-03-01

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

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

```
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_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_12_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? 9

Q6. Note that there are "missing values" in the dataset. How many NA values there in the persons_fully_vaccinated column? 18338

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

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

Q7. What percent of persons_fully_vaccinated values are missing (to 2 significant figures)? 17.04%

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

```
## [1] 17.04212
```

Q8. [Optional]: Why might this data be missing? People might have not wanted to answer if they were vaccinated or not.

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)
```

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

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

```
age <- today()-ymd("2022-03-03")
age
```

```
## Time difference of 0 days
```

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

```
## [1] 0
```

First I have to make sure my covid vaccination data date column is in lubridate format.

```
# 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? 422

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

```
## Time difference of 422 days
```

Q10. How many unique dates are in the dataset (i.e. how many different dates are detailed)? 61

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

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

Working with ZIP codes

```
library(zipcodeR)
geocode_zip('92037')
```

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

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

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

```
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>
```

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:

We have done this the *base R* way quite often like so

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

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

An often more convinent 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
```

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

Q11. How many distinct zip codes are listed for San Diego County? 107

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

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

Q12. What San Diego County Zip code area has the largest 12 + Population in this dataset? 92154

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

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

Using dplyr select all San Diego "county" entries on "as_of_date" "2022-03-01" and use this for the following questions.

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

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

Let's do this with the most recent date in the data-set (2022-03-01).

Q13. What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2022-03-01"? 0.705

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

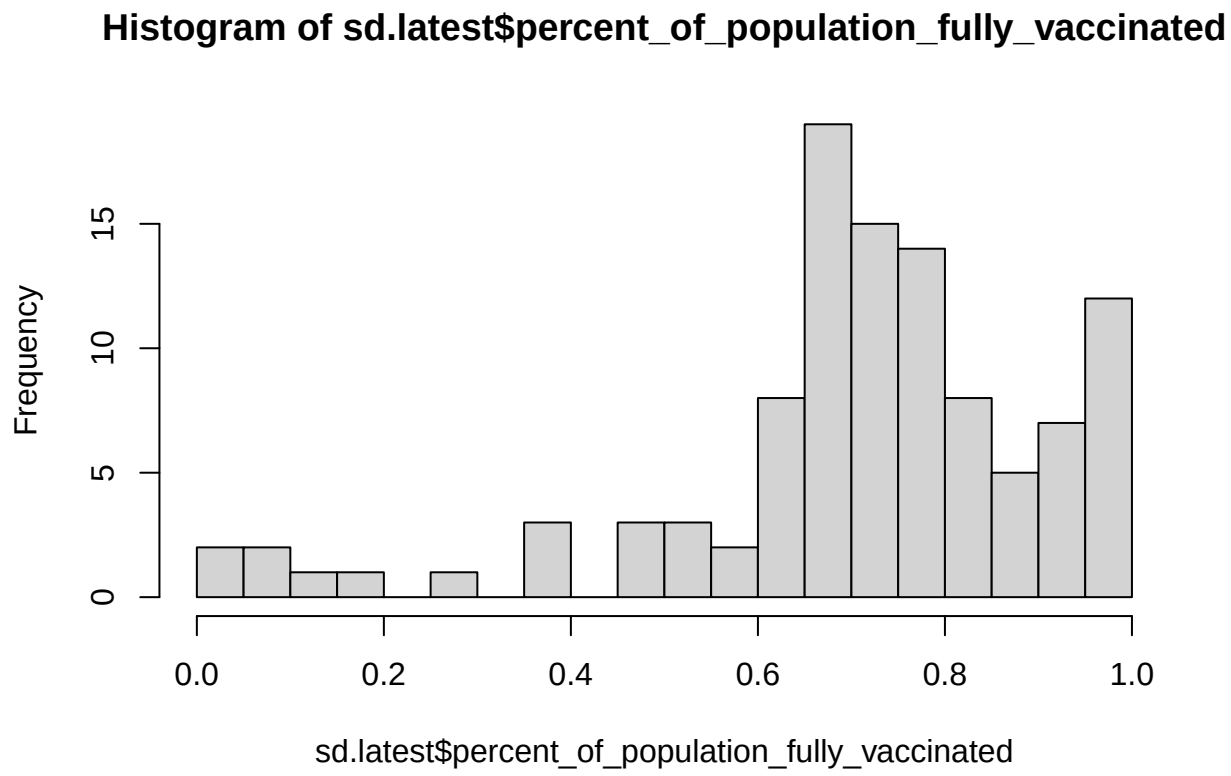
```
## [1] 0.7052904
```

```
summary(sd.latest$percent_of_population_fully_vaccinated, na.rm=TRUE)
```

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

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 "2022-03-01"?

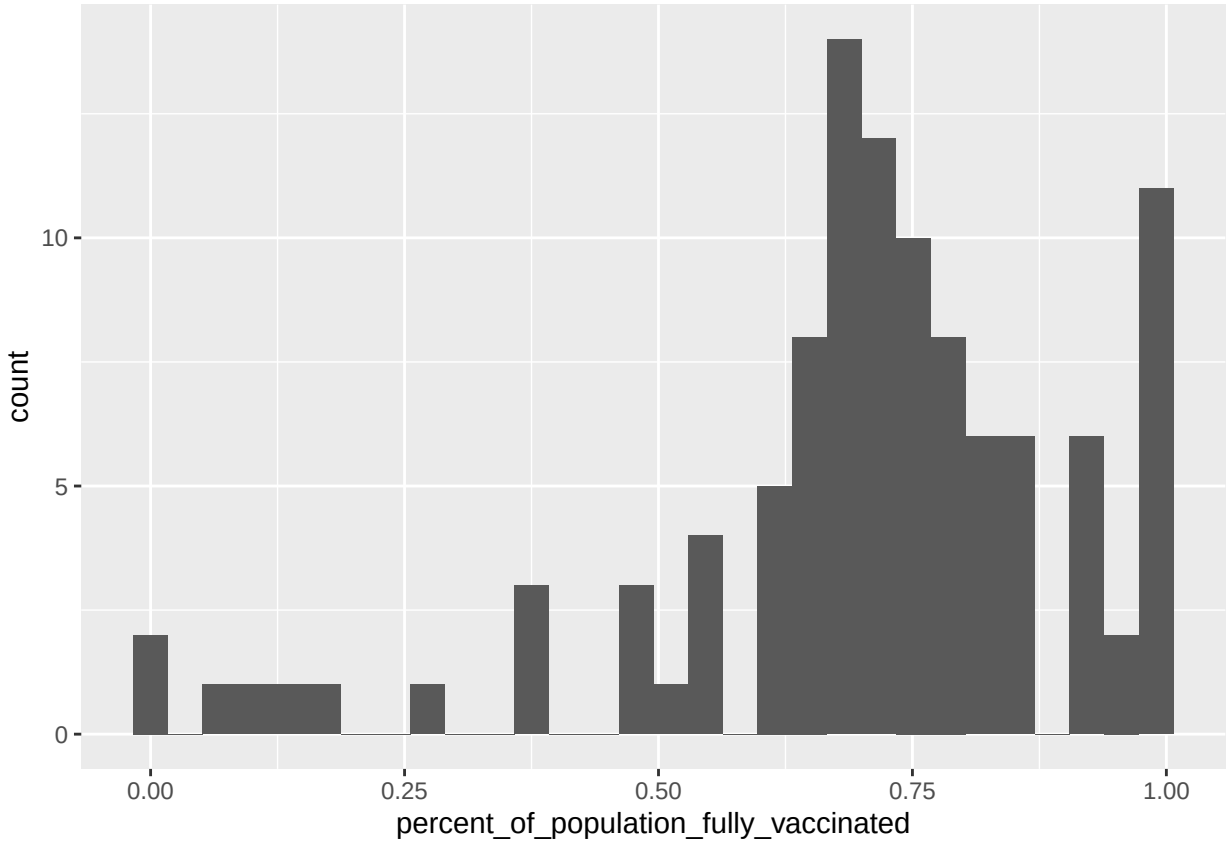
```
hist(sd.latest$percent_of_population_fully_vaccinated, breaks=30)
```




```
library(ggplot2)
ggplot(sd.latest) +
  aes(percent_of_population_fully_vaccinated) +
  geom_histogram()
```

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

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



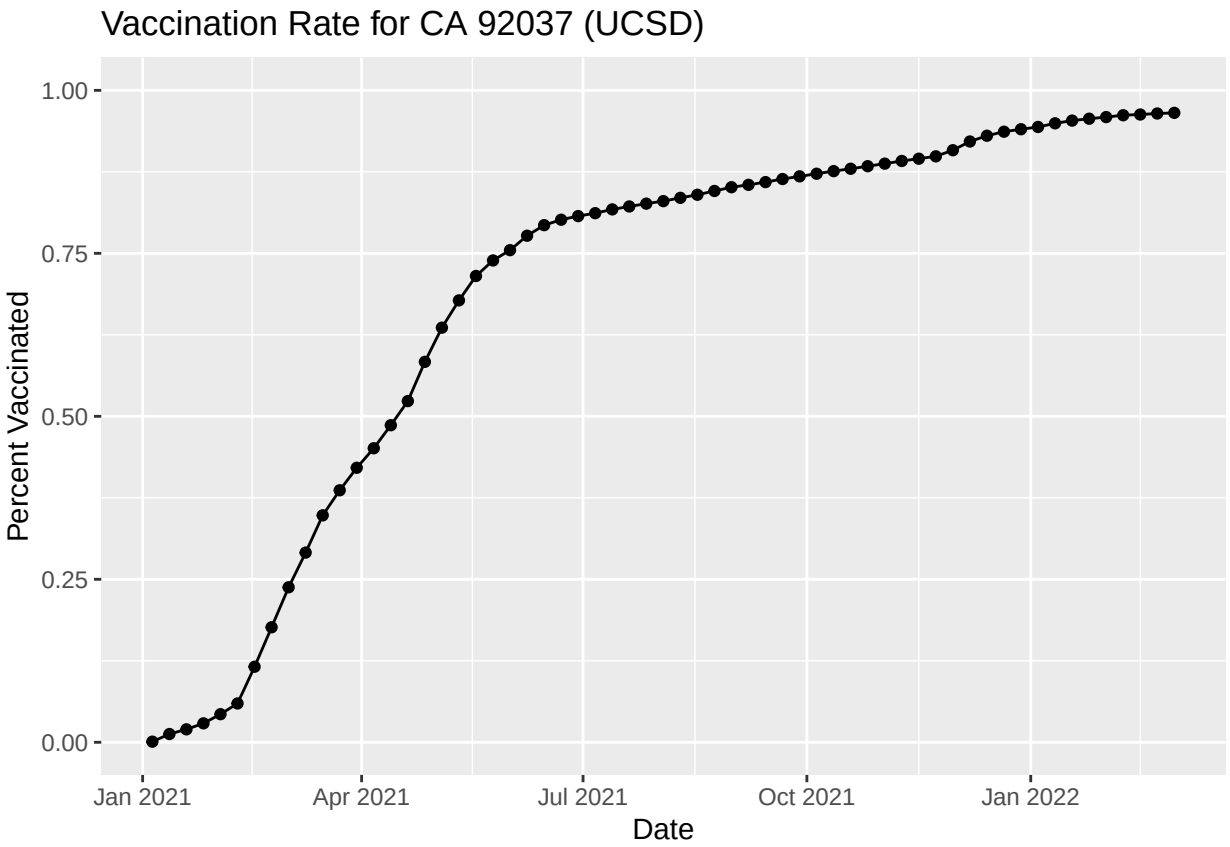
Focus on UCSD/La Jolla

```
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:

```
baseplot <- 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 CA 92037 (UCSD)")
baseplot
```



Comparing to similar sized areas

```
# Subset to all CA areas with a population as large as 92037
vax.36 <- filter(vax, age5_plus_population > 36144 &
                 as_of_date == "2022-03-01")
head(vax.36)
```

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction      county
## 1 2022-03-01                   95628                Sacramento  Sacramento
## 2 2022-03-01                   90808                Long Beach Los Angeles
## 3 2022-03-01                   92507                 Riverside   Riverside
## 4 2022-03-01                   92626                    Orange      Orange
## 5 2022-03-01                   93257                    Tulare      Tulare
## 6 2022-03-01                   90011               Los Angeles Los Angeles
##   vaccine_equity_metric_quartile                vem_source
## 1                              3 Healthy Places Index Score
## 2                              4 Healthy Places Index Score
## 3                              1 Healthy Places Index Score
## 4                              3 Healthy Places Index Score
## 5                              1 Healthy Places Index Score
## 6                              1 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1               35579.0                38694                    28842
## 2               33952.3                37179                    29383
```

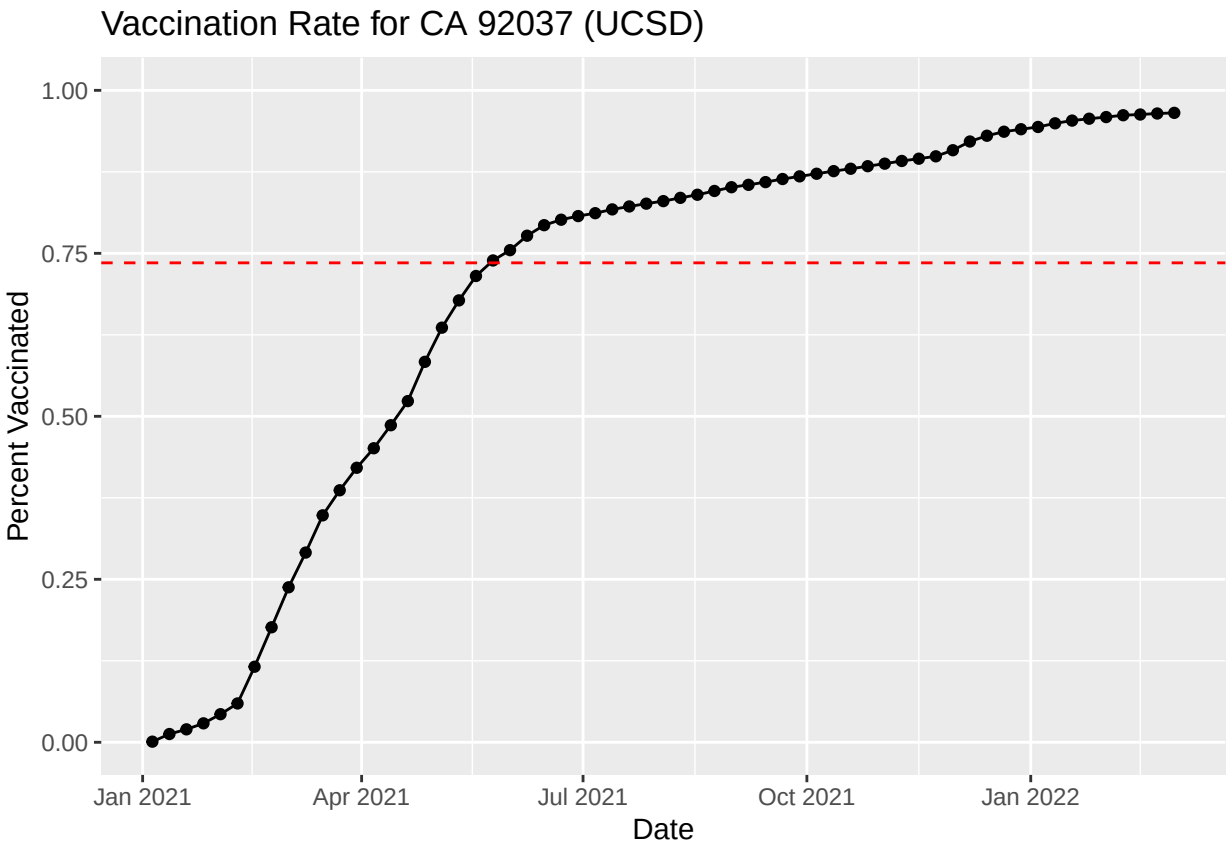
```
## 3               51432.5               55253                    34455
## 4               44238.8               47883                    33767
## 5               61519.8               70784                    42919
## 6               87902.8              101902                    65342
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                         1990                                0.745387
## 2                         2112                                0.790312
## 3                         3947                                0.623586
## 4                         2937                                0.705198
## 5                         5868                                0.606338
## 6                        15255                                0.641224
##   percent_of_population_partially_vaccinated
## 1                                   0.051429
## 2                                   0.056806
## 3                                   0.071435
## 4                                   0.061337
## 5                                   0.082900
## 6                                   0.149703
##   percent_of_population_with_1_plus_dose booster_recip_count redacted
## 1                               0.796816               16913       No
## 2                               0.847118               17253       No
## 3                               0.695021               15073       No
## 4                               0.766535               17595       No
## 5                               0.689238               17740       No
## 6                               0.790927               19928       No
```

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

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

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?

```
baseplot +
  geom_hline(yintercept = ave.36, 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, na.rm=TRUE)
```

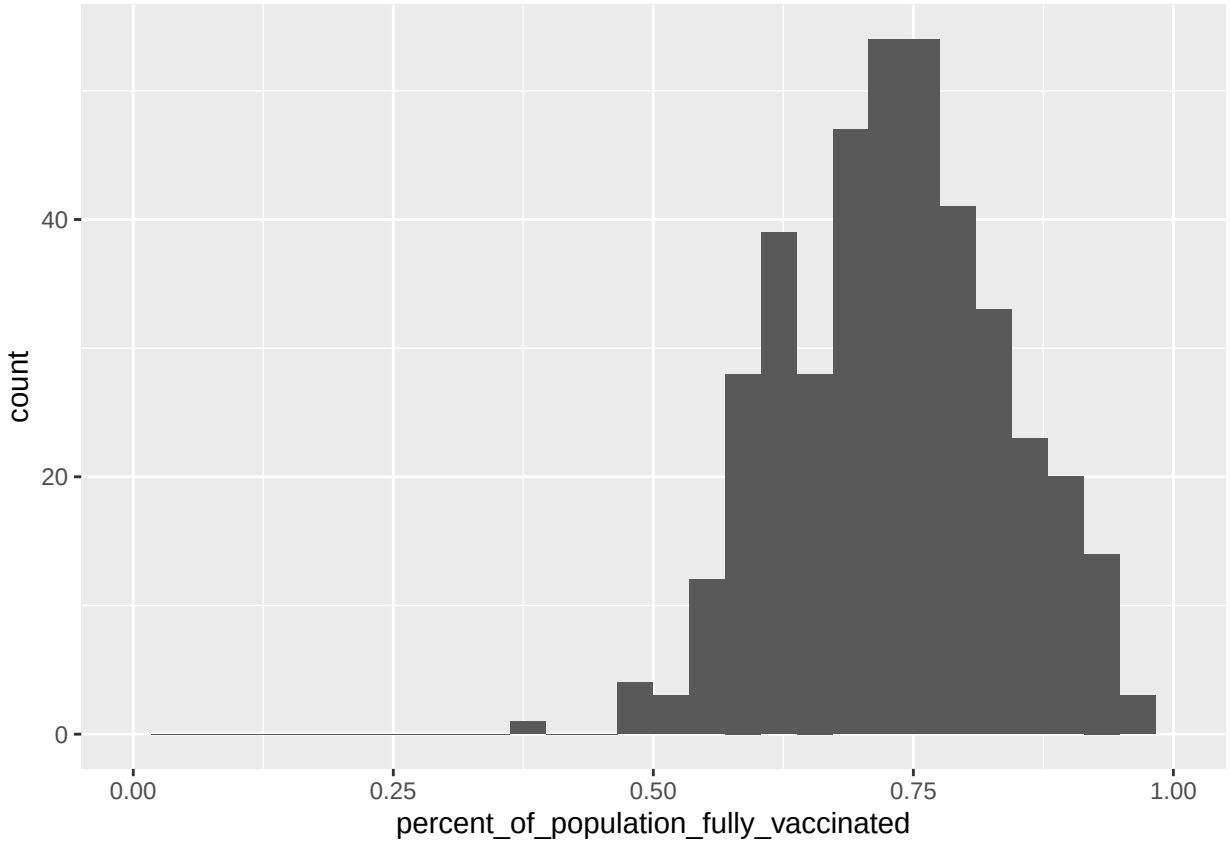
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3890  0.6554  0.7350  0.7354  0.8044  1.0000
```

Q18. Using ggplot generate a histogram of this data.

```
ggplot(vax.36) +
  aes(percent_of_population_fully_vaccinated) +
  geom_histogram() +
  xlim(c(0,1))
```

```
## `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? Both the 92109 and 92040 ZIP codes are below the average.

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

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

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

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

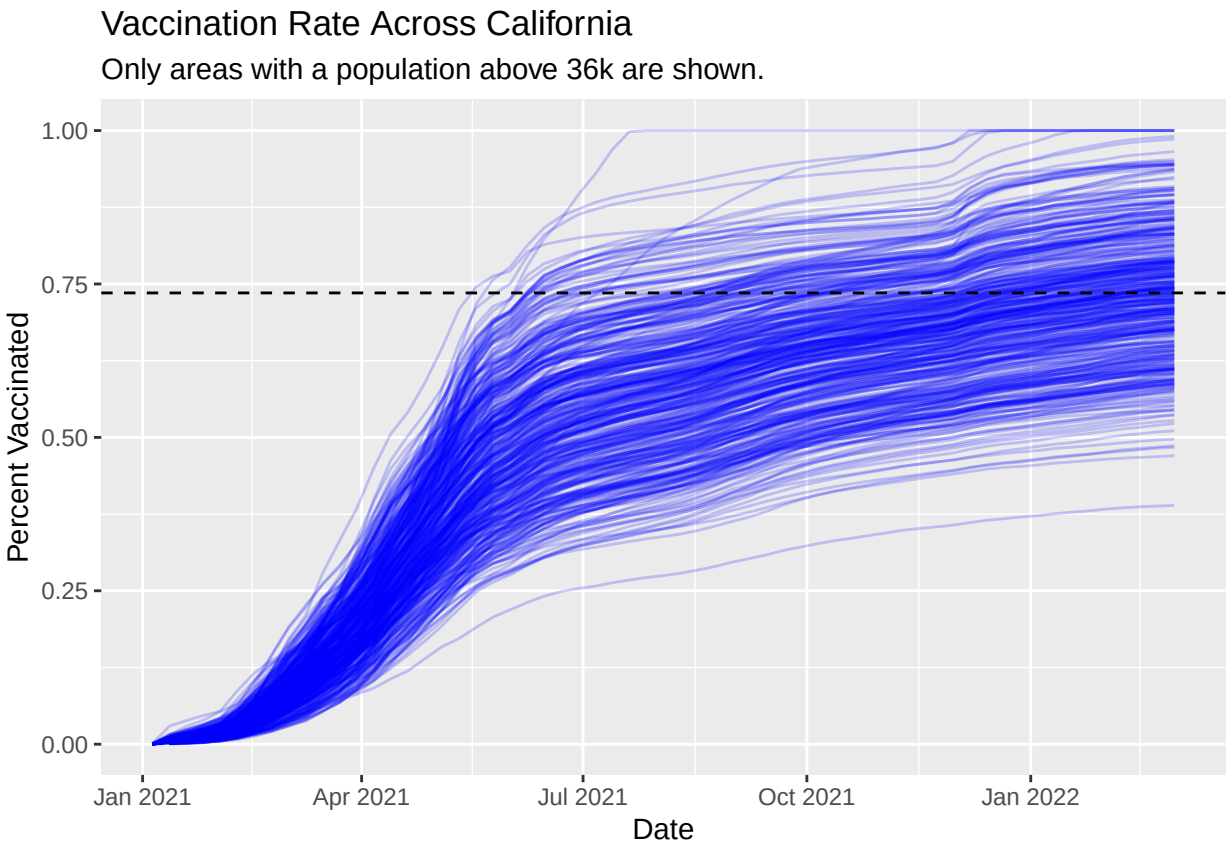
Q20. Finally make a time course plot of vaccination progress for all areas in the full dataset with a age5_plus_population $>$ 36144.

```
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 = ave.36, linetype=2)
```

```
## 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 little nervous but okay with it.

```
sessionInfo()
```

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib
```

```
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_3.3.5   dplyr_1.0.8     zipcodeR_0.3.3  lubridate_1.8.0
##
## loaded via a namespace (and not attached):
##  [1] httr_1.4.2         tidyr_1.2.0        bit64_4.0.5        jsonlite_1.7.3
##  [5] sp_1.4-6           highr_0.9          blob_1.2.2         yaml_2.2.2
##  [9] tidycensus_1.1     pillar_1.7.0       RSQLite_2.2.10     lattice_0.20-45
## [13] glue_1.6.1         uuid_1.0-3         digest_0.6.29      rvest_1.0.2
## [17] colorspace_2.0-2   htmltools_0.5.2    pkgconfig_2.0.3    raster_3.5-15
## [21] purrr_0.3.4        scales_1.1.1       terra_1.5-21       tzdb_0.2.0
## [25] tigris_1.6         tibble_3.1.6       proxy_0.4-26       farver_2.1.0
## [29] generics_0.1.2     ellipsis_0.3.2     cachem_1.0.6       withr_2.4.3
## [33] repr_1.1.4         skimr_2.1.3        cli_3.1.1          magrittr_2.0.2
## [37] crayon_1.4.2       memoise_2.0.1      maptools_1.1-2     evaluate_0.14
## [41] fansi_1.0.2        xml2_1.3.3         foreign_0.8-81     class_7.3-19
## [45] tools_4.1.2        hms_1.1.1          lifecycle_1.0.1    stringr_1.4.0
## [49] munsell_0.5.0      compiler_4.1.2     e1071_1.7-9        rlang_1.0.0
## [53] classInt_0.4-3     units_0.8-0        grid_4.1.2         rstudioapi_0.13
## [57] rappdirs_0.3.3     labeling_0.4.2     base64enc_0.1-3    rmarkdown_2.11
## [61] gtable_0.3.0       codetools_0.2-18   DBI_1.1.2          curl_4.3.2
## [65] R6_2.5.1           knitr_1.37         rgdal_1.5-28       fastmap_1.1.0
## [69] bit_4.0.4          utf8_1.2.2         KernSmooth_2.23-20 readr_2.1.2
## [73] stringi_1.7.6      Rcpp_1.0.8         vctrs_0.3.8        sf_1.0-6
## [77] tidyselect_1.1.1   xfun_0.29
```