class 17: COVID Mini Project

Jimmi

Be sure to move your downloaded CSV file to your project directory and then read/import into an R object called vax. We will 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                   95446                     Sonoma    Sonoma
2 2021-01-05                   96014                   Siskiyou  Siskiyou
3 2021-01-05                   96087                     Shasta    Shasta
4 2021-01-05                   96008                     Shasta    Shasta
5 2021-01-05                   95410                  Mendocino Mendocino
6 2021-01-05                   95527                    Trinity   Trinity
  vaccine_equity_metric_quartile                vem_source
1                              2 Healthy Places Index Score
2                              2    CDPH-Derived ZCTA Score
3                              2    CDPH-Derived ZCTA Score
4                             NA            No VEM Assigned
5                              3    CDPH-Derived ZCTA Score
6                              2    CDPH-Derived ZCTA Score
  age12_plus_population age5_plus_population tot_population
1                4840.7                 5057           5168
2                 135.0                  135            135
3                 513.9                  544            544
4                1125.3                 1164             NA
5                 926.3                  988            997
6                 476.6                  485            499
  persons_fully_vaccinated persons_partially_vaccinated
1                       NA                           NA
2                       NA                           NA
```

```
3                       NA                            NA
4                       NA                            NA
5                       NA                            NA
6                       NA                            NA
  percent_of_population_fully_vaccinated
1                                     NA
2                                     NA
3                                     NA
4                                     NA
5                                     NA
6                                     NA
  percent_of_population_partially_vaccinated
1                                         NA
2                                         NA
3                                         NA
4                                         NA
5                                         NA
6                                         NA
  percent_of_population_with_1_plus_dose booster_recip_count
1                                     NA                  NA
2                                     NA                  NA
3                                     NA                  NA
4                                     NA                  NA
5                                     NA                  NA
6                                     NA                  NA
  bivalent_dose_recip_count eligible_recipient_count
1                        NA                        0
2                        NA                        0
3                        NA                        2
4                        NA                        2
5                        NA                        0
6                        NA                        0
                                                                    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
```

```
attributes(vax)
```

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

persons_fully_vaccinated details peoples that are fully vaccinated. Shown through `attributes` function.

Q2. What column details the Zip code tabulation area?

zip_code_tabulation_area details zip code tabulation areas. Shown through `attributes` function.

Q3. What is the earliest date in this dataset?

```
min(vax$as_of_date)
```

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

Q4. What is the latest date in this dataset?

```
max(vax$as_of_date)
```

```
[1] "2023-02-28"
```

As we have done previously, let's call the `skim()` function from the skimr package to get a quick overview of this dataset:

```
library(skimr)
skimmed = skim(vax)
skimmed
```

Table 1: Data summary

Name	vax
Number of rows	199332
Number of columns	18
_____	
Column type frequency:	
character	5
numeric	13
_____	
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	113	0
local_health_jurisdiction	0	1	0	15	565	62	0
county	0	1	0	15	565	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	1817.38	90001	92257.75	93658.50	95380.50	97635.0	
vaccine_equity_metric_quartile	9831	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	
age12_plus_population	0	1.00	18895.01	18993.87	0	1346.95	13685.13	31756.18	88556.7	
age5_plus_population	0	1.00	20875.22	21105.97	0	1460.50	15364.00	34877.00	101902.0	
tot_population	9718	0.95	23372.72	22628.51	12	2126.00	18714.00	38168.00	111165.0	
persons_fully_vaccinated	16525	0.92	13962.33	15054.09	11	930.00	8566.00	23302.00	87566.0	
persons_partially_vaccinated	16525	0.92	1701.64	2030.18	11	165.00	1196.00	2535.00	39913.0	
percent_of_population_fully_vaccinated	20825	0.90	0.57	0.25	0	0.42	0.60	0.74	1.0	
percent_of_population_partially_vaccinated	20825	0.90	0.08	0.09	0	0.05	0.06	0.08	1.0	
percent_of_population_with_1_plus_dose	21859	0.89	0.63	0.24	0	0.49	0.67	0.81	1.0	
booster_recip_count	72872	0.63	5837.31	7165.81	11	297.00	2748.00	9438.25	59553.0	
bivalent_dose_recip_count	158664	0.20	2924.93	3583.45	11	190.00	1418.00	4626.25	27458.0	
eligible_recipient_count	0	1.00	12801.84	14908.33	0	504.00	6338.00	21973.00	67234.0	

Q5. How many numeric columns are in this dataset?

There are 13 numeric columns.

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

There are 16525 NA values.

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

total persons_fully_vaccinated = 87566 and the missing NA values = 16525

```
signif(16525/87566*100, 4)
```

```
[1] 18.87
```

percent that are missing is 18.87%

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

```
library(lubridate)
```

```
Attaching package: 'lubridate'

The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

```
today()
```

```
[1] "2023-03-13"
```

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 how many days have passed since data was first recorded

```
# 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 797 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 784 days
```

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

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

```
Time difference of 13 days
```

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

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

```
[1] 113
```

```
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
```

```
n_distinct(vax$as_of_date)
```

```
[1] 113
```

```
library(zipcodeR)

geocode_zip('92108')
```

```
# A tibble: 1 x 3
  zipcode   lat   lng
  <chr>   <dbl> <dbl>
1 92108    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_~1 major~2 post_~3 common_c~4 county state   lat   lng timez~5
  <chr>   <chr>      <chr>   <chr>       <blob> <chr>  <chr> <dbl> <dbl> <chr>
1 92037   Standard   La Jol~ La Jol~ <raw 20 B> San D~ CA     32.8 -117. Pacific
2 92109   Standard   San Di~ San Di~ <raw 21 B> San D~ CA     32.8 -117. Pacific
# ... with 14 more variables: 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>, and abbreviated variable names
#   1: zipcode_type, 2: major_city, 3: post_office_city, ...
```

```
reverse_zipcode(c('92037',"92109"))
```

```
# A tibble: 2 x 24
  zipcode zipcode_~1 major~2 post_~3 common_c~4 county state   lat   lng timez~5
  <chr>   <chr>      <chr>   <chr>       <blob> <chr>  <chr> <dbl> <dbl> <chr>
1 92037   Standard   La Jol~ La Jol~ <raw 20 B> San D~ CA     32.8 -117. Pacific
2 92109   Standard   San Di~ San Di~ <raw 21 B> San D~ CA     32.8 -117. Pacific
# ... with 14 more variables: 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>, and abbreviated variable names
#   1: zipcode_type, 2: major_city, 3: post_office_city, ...
```

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" , ]

nrow(sd)
```

```
[1] 12091
```

It is time to revisit the most awesome **dplyr** package.

```
library(dplyr)
```

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

nrow(sd.10)
```

```
[1] 8588
```

How many ZIP codes are we dealing with?

```
n_distinct(sd.10$zip_code_tabulation_area)
```

```
[1] 76
```

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

```
length(unique(sd))
```

```
[1] 18
```

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

```
ind = which.max(sd$age12_plus_population)

sd$zip_code_tabulation_area[2]
```

```
[1] 92154
```

```
reverse_zipcode("92154")
```

```
# A tibble: 1 x 24
  zipcode zipcode_~1 major~2 post_~3 common_c~4 county state   lat   lng timez~5
  <chr>   <chr>      <chr>   <chr>   <blob>     <chr>  <chr> <dbl> <dbl> <chr>
1 92154   Standard   San Di~ San Di~ <raw 21 B> San D~ CA     32.6  -117 Pacific
# ... with 14 more variables: 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>, and abbreviated variable names
#   1: zipcode_type, 2: major_city, 3: post_office_city, ...
```

Q13. What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2022-11-15"?

```
sd.b4 <- filter(vax, county == "San Diego" &
                as_of_date == "2022-11-15")

head(sd.b4)
```

```
  as_of_date zip_code_tabulation_area local_health_jurisdiction    county
1 2022-11-15                   92124                San Diego San Diego
2 2022-11-15                   91901                San Diego San Diego
3 2022-11-15                   91902                San Diego San Diego
4 2022-11-15                   92064                San Diego San Diego
5 2022-11-15                   92069                San Diego San Diego
6 2022-11-15                   92009                San Diego San Diego
  vaccine_equity_metric_quartile                vem_source
1                              3 Healthy Places Index Score
2                              3 Healthy Places Index Score
3                              4 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 tot_population
1               25422.4                29040          32600
2               15549.8                16905          18162
3               16620.7                18026          18896
4               42177.1                46855          49805
```

```
5               41447.3                46850          50376
6               39183.5                43710          46612
  persons_fully_vaccinated persons_partially_vaccinated
1                    18753                         2304
2                     9764                          727
3                    14906                         1670
4                    36984                         2713
5                    34945                         2795
6                    34282                         2647
  percent_of_population_fully_vaccinated
1                               0.575245
2                               0.537606
3                               0.788844
4                               0.742576
5                               0.693684
6                               0.735476
  percent_of_population_partially_vaccinated
1                                   0.070675
2                                   0.040029
3                                   0.088378
4                                   0.054472
5                                   0.055483
6                                   0.056788
  percent_of_population_with_1_plus_dose booster_recip_count
1                               0.645920               10460
2                               0.577635                4849
3                               0.877222                8164
4                               0.797048               20769
5                               0.749167               17753
6                               0.792264               20238
  bivalent_dose_recip_count eligible_recipient_count redacted
1                      3931                    18663       No
2                      1286                     9754       No
3                      2285                    14873       No
4                      7002                    36884       No
5                      4255                    34882       No
6                      7708                    34215       No
```

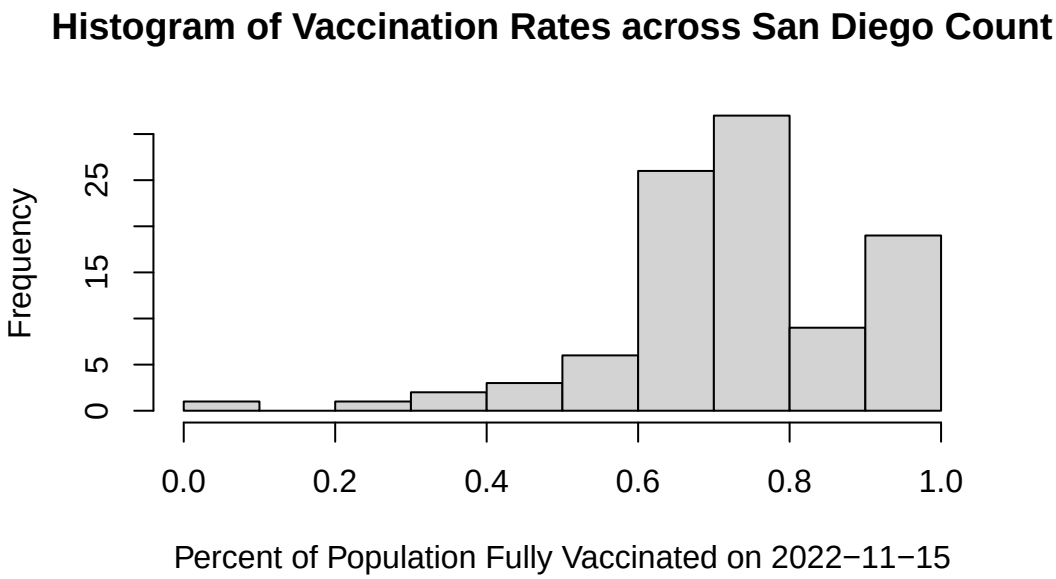
```
mean(sd.b4$percent_of_population_fully_vaccinated, na.rm=T)
```

```
[1] 0.7380708
```

The average percent of population fully vaccinated as of 2022-11-15 is 0.7380708

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 "2022-11-15"?

```
hist(sd.b4$percent_of_population_fully_vaccinated,
     main="Histogram of Vaccination Rates across San Diego County",
      xlab="Percent of Population Fully Vaccinated on 2022-11-15",
     )
```



Focus on UCSD/La Jolla

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

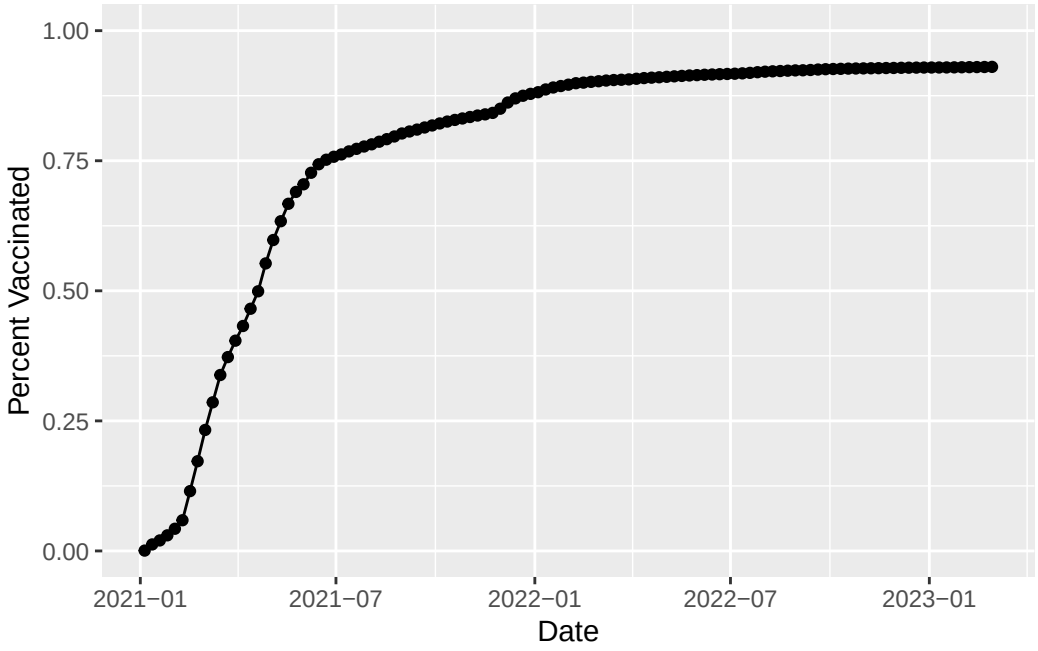
```
[1] 36144
```

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

```
library(ggplot2)
```

```
gg = 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")

gg
```



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 == "2022-11-15")
```

```
#head(vax.36)
```

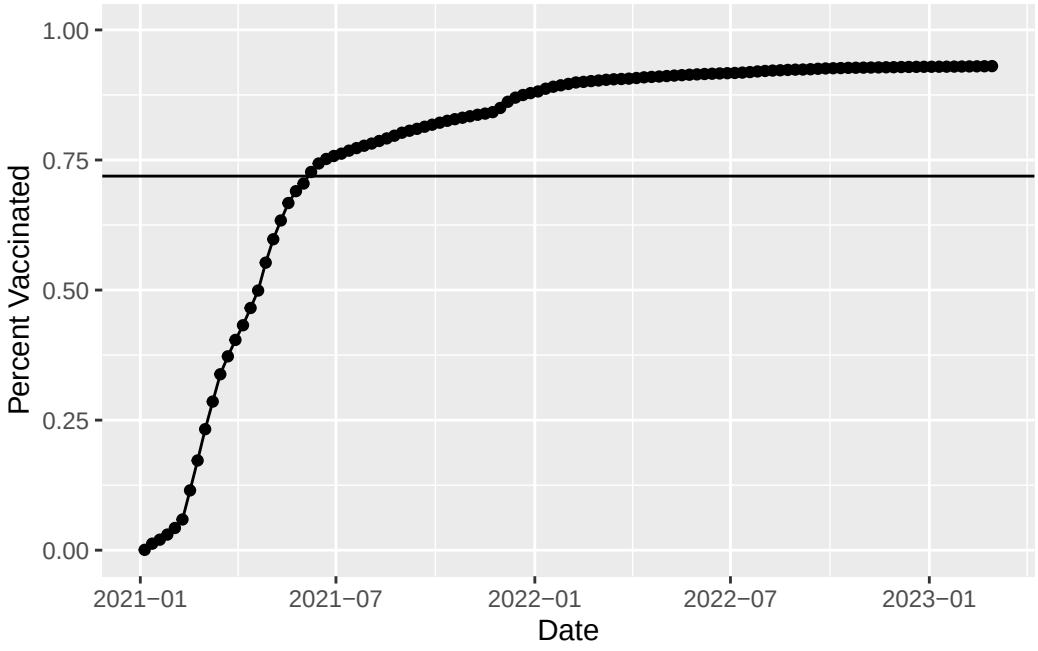
Q16. Calculate the mean "Percent of Population Fully Vaccinated" for ZIP code areas with a population as large as 92037 (La Jolla) as_of_date "2022-11-15". Add this as a straight horizontal line to your plot from above with the geom_hline() function?

```
ave = mean(vax.36$percent_of_population_fully_vaccinated)
ave
```

```
[1] 0.7190515
```

The mean was 0.7190515% which will be added as a line to the ggplot.

```
gg + geom_hline(yintercept = ave)
```



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 "2022-11-15"?

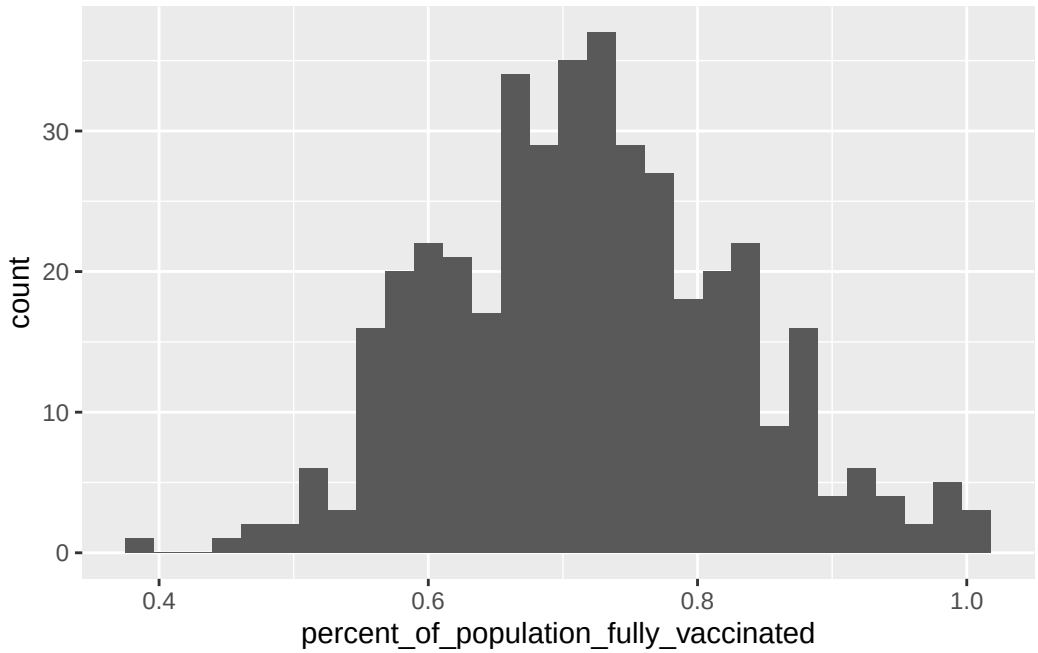
```
summary(vax.36$percent_of_population_fully_vaccinated)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.3784  0.6444  0.7162  0.7191  0.7882  1.0000 
```

Q18. Using ggplot generate a histogram of this data.

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

```
`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?

```
x = filter(vax.36, zip_code_tabulation_area %in% c("92109", "92040"))
x$percent_of_population_fully_vaccinated
```

```
[1] 0.548849 0.692874
```

```
ave>x$percent_of_population_fully_vaccinated
```

```
[1] TRUE TRUE
```

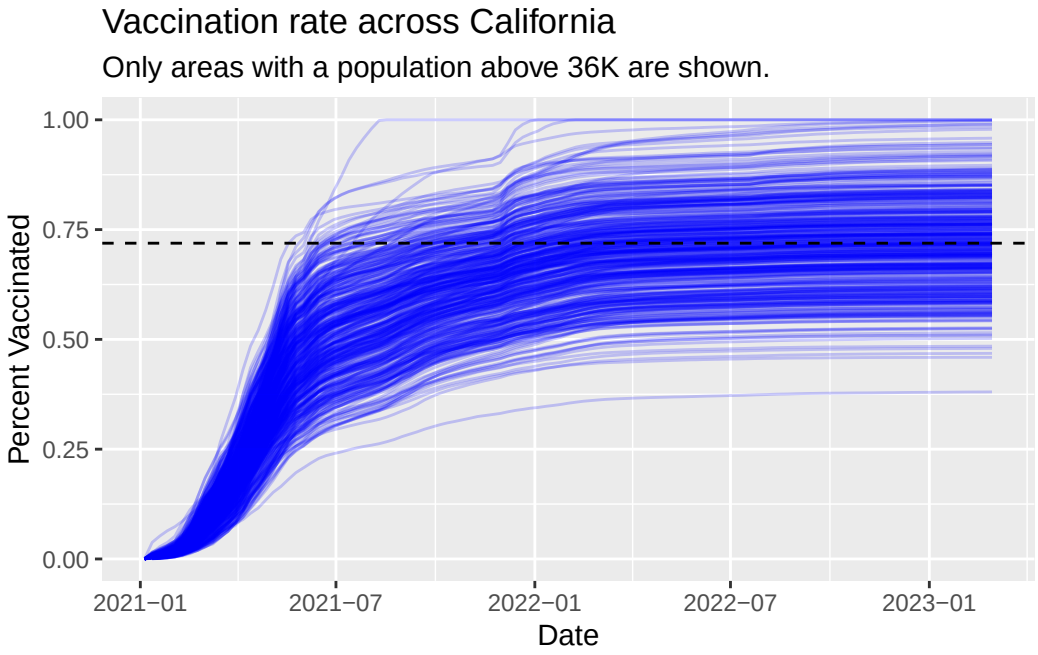
They are both below the average.

> 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 areas with a population above 36K are shown.") +
  geom_hline(yintercept = ave, linetype="dashed")
```

```
Warning: Removed 183 rows containing missing values (`geom_line()`).
```




```
sessionInfo()
```

```
R version 4.2.2 (2022-10-31 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 22621)

Matrix products: default

locale:
[1] LC_COLLATE=English_United States.utf8
[2] LC_CTYPE=English_United States.utf8
[3] LC_MONETARY=English_United States.utf8
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.utf8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] ggplot2_3.4.1   zipcodeR_0.3.5  dplyr_1.1.0     lubridate_1.9.2
[5] skimr_2.1.5
```

```
loaded via a namespace (and not attached):
 [1] Rcpp_1.0.10         lattice_0.20-45     tidyr_1.3.0         class_7.3-20
 [5] digest_0.6.31       utf8_1.2.3          R6_2.5.1            repr_1.1.6
 [9] RSQLite_2.3.0       evaluate_0.20       e1071_1.7-13        httr_1.4.5
[13] pillar_1.8.1        rlang_1.0.6         curl_5.0.0          uuid_1.1-0
[17] rstudioapi_0.14     raster_3.6-20       blob_1.2.3          rmarkdown_2.20
[21] labeling_0.4.2      readr_2.1.4         stringr_1.5.0       munsell_0.5.0
[25] bit_4.0.5           proxy_0.4-27        compiler_4.2.2      xfun_0.37
[29] pkgconfig_2.0.3     tigris_2.0.1        base64enc_0.1-3     htmltools_0.5.4
[33] tidyselect_1.2.0    tibble_3.1.8        codetools_0.2-18    fansi_1.0.4
[37] crayon_1.5.2        tzdb_0.3.0          withr_2.5.0         sf_1.0-9
[41] tidycensus_1.3.2    rappdirs_0.3.3      grid_4.2.2          gtable_0.3.1
[45] jsonlite_1.8.4      lifecycle_1.0.3     DBI_1.1.3           magrittr_2.0.3
[49] scales_1.2.1        units_0.8-1         KernSmooth_2.23-20  cli_3.6.0
[53] stringi_1.7.12      cachem_1.0.7        farver_2.1.1        sp_1.6-0
[57] xml2_1.3.3          ellipsis_0.3.2      generics_0.1.3      vctrs_0.5.2
[61] tools_4.2.2         bit64_4.0.5         glue_1.6.2          purrr_1.0.1
[65] hms_1.1.2           fastmap_1.1.1       yaml_2.3.7          colorspace_2.1-0
[69] timechange_0.2.0    terra_1.7-18        classInt_0.4-9      rvest_1.0.3
[73] memoise_2.0.1       knitr_1.42
```