Class17

Gregory Jordan

Getting Started

```
#import vaccination data
vax <- read.csv("https://data.chhs.ca.gov/dataset/ead44d40-fd63-4f9f-950a-3b0111074de8/res
head(vax)
```

```
  as_of_date zip_code_tabulation_area local_health_jurisdiction          county
1 2021-01-05                   92240                 Riverside       Riverside
2 2021-01-05                   91302               Los Angeles     Los Angeles
3 2021-01-05                   93420           San Luis Obispo San Luis Obispo
4 2021-01-05                   91901                 San Diego       San Diego
5 2021-01-05                   94110             San Francisco   San Francisco
6 2021-01-05                   91902                 San Diego       San Diego
  vaccine_equity_metric_quartile                 vem_source
1                              1 Healthy Places Index Score
2                              4 Healthy Places Index Score
3                              3 Healthy Places Index Score
4                              3 Healthy Places Index Score
5                              4 Healthy Places Index Score
6                              4 Healthy Places Index Score
  age12_plus_population age5_plus_population tot_population
1               29270.5                33093          35278
2               23163.9                25899          26712
3               26694.9                29253          30740
4               15549.8                16905          18162
5               64350.7                68320          72380
6               16620.7                18026          18896
  persons_fully_vaccinated persons_partially_vaccinated
1                       NA                           NA
2                       15                          614
```

```
3                       NA                           NA
4                       NA                           NA
5                       17                         1268
6                       15                          397
  percent_of_population_fully_vaccinated
1                                     NA
2                                0.000562
3                                     NA
4                                     NA
5                                0.000235
6                                0.000794
  percent_of_population_partially_vaccinated
1                                         NA
2                                    0.022986
3                                         NA
4                                         NA
5                                    0.017519
6                                    0.021010
  percent_of_population_with_1_plus_dose booster_recip_count
1                                     NA                  NA
2                                0.023548                 NA
3                                     NA                  NA
4                                     NA                  NA
5                                0.017754                 NA
6                                0.021804                 NA
  bivalent_dose_recip_count eligible_recipient_count
1                        NA                       2
2                        NA                      15
3                        NA                       4
4                        NA                       8
5                        NA                      17
6                        NA                      15
                                                                    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
```

Note for the project: The dataset I got from the website is different than Barry's because the dataset has been updated since Barry made the assignment. and I could not find Barry's no

matter how hard I tried, so I will follow the exercises and answer the questions according to my dataset, but I will also include the values for Barry's dataset in the questions to show I know what the correct answers are.

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?

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

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

Q4. What is the latest date in this dataset?

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

```
[1] "2022-11-22"
```

For Barry's dataset the latest date is 2022-11-15

Use Skimr to skim the dataset

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

Table 1: Data summary

Name	vax
Number of rows	174636
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	99	0
local_health_jurisdiction	0	1	0	15	495	62	0
county	0	1	0	15	495	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	8613	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	
age12_plus_population	0	1.00	18895.01	18993.88	0	1346.95	13685.13	31756.12	88556.7	
age5_plus_population	0	1.00	20875.22	21105.98	0	1460.50	15364.00	34877.00	101902.0	
tot_population	8514	0.95	23372.72	22628.51	12	2126.00	18714.00	38168.00	111165.0	
persons_fully_vaccinated	14921	0.91	13466.31	14722.46	11	883.00	8024.00	22529.00	67186.0	
persons_partially_vaccinated	14921	0.91	1707.50	1998.80	11	167.00	1194.00	2547.00	39204.0	
percent_of_population_fully_vaccinated	18665	0.89	0.55	0.25	0	0.39	0.59	0.73	1.0	
percent_of_population_partially_vaccinated	18665	0.89	0.08	0.09	0	0.05	0.06	0.08	1.0	
percent_of_population_with_1_plus_dose	19362	0.89	0.61	0.25	0	0.46	0.65	0.79	1.0	
booster_recip_count	70421	0.60	5655.17	6867.49	11	280.00	2575.00	9421.00	58304.0	
bivalent_dose_recip_count	156958	0.10	1646.02	2161.84	11	109.00	719.00	2443.00	18109.0	
eligible_recipient_count	0	1.00	12309.11	14555.83	0	466.00	5810.00	21140.00	86696.0	

Q5. How many numeric columns are in this dataset?

13

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

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

```
[1] 14921
```

Note: My dataset has 14921, but Barry's in the webpage has 15440.

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

```
cat(round(sum(is.na(vax$persons_fully_vaccinated))/(sum(is.na(vax$persons_fully_vaccinated
```

```
8.54 %
```

Note: Again, because my dataset is different I got 8.54% wheras Barry's should be 8.93% (15440 persons fully vaccinated / 172872 total rows * 100%)

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

It is likely due to privacy laws as evident by the "information redacted" section in the dataset. It could also be due to bad reporteing or difficulty finding the data to add it to the dataset.

Working with Dates

use `lubridate` package to make life better when working with datetime in R

```
#load in lubridate package
library(lubridate)
```

```
Loading required package: timechange
```

```

Attaching package: 'lubridate'
```

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

    date, intersect, setdiff, union
```

```
#what is today's date
today()
```

```
[1] "2022-11-23"
```

```
#we will get an error if we use vax$as_of_date as it currently is becauase it needs to be
today() - vax$as_of_date[1]
```

```
Error in unclass(as.Date(e1)) - e2: non-numeric argument to binary operator
```

```
#convert as_of_date to lubridate format using ymd (year month day) format
vax$as_of_date <- ymd(vax$as_of_date)
```

now we can do math with the dates because we have transformed our vax$as_of_date

```
#how many days have passed since the first vax report in the dataset
today()-vax$as_of_date[1]
```

```
Time difference of 687 days
```

```
#how many days does the dataset span?
vax$as_of_date[nrow(vax)]-vax$as_of_date[1]
```

```
Time difference of 686 days
```

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

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

```
Time difference of 1 days
```

Note: For Barry's dataset the time difference is 6 days

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

```
nrow(as.data.frame(unique(vax$as_of_date)))
```

```
[1] 99
```

Note: For Barry's dataset he used in the assignment there are 98 unique dates because his dataset is less recent

Working with Zip Codes

use `zipcodeR` package to work with zip codes

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

```
#get lat and long of la jolla zip code
geocode_zip("92037")
```

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

```
#get distance between centroids of zip codes
zip_distance("92037","92109")
```

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

Focus on the San Diego Area

```
#subset vax to san diego
#i like 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")

nrow(sd)
```

```
[1] 10593
```

using **dplyr** often more convenient when subsetting over multiple variables

ex. all san diego counties with population over 10,000

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

```
[1] 7524
```

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

```
nrow(as.data.frame(unique(sd$zip_code_tabulation_area)))
```

```
[1] 107
```

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

```
#using dplyr
head(filter(vax,county=="San Diego") %>% arrange(desc(age12_plus_population)),1)[2]
```

```
  zip_code_tabulation_area
1                    92154
```

select all san diego county entries on as of date 2022-11-15

```
sd.2022.11.15 <- filter(sd,as_of_date=="2022-11-15")
```

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

```
#get the average percent of population fully vaccinated for as of date = 2022-11-15
#remember to remove na values
```

```
cat(round(mean(sd.2022.11.15$percent_of_population_fully_vaccinated,na.rm = TRUE)*100,2),"
```

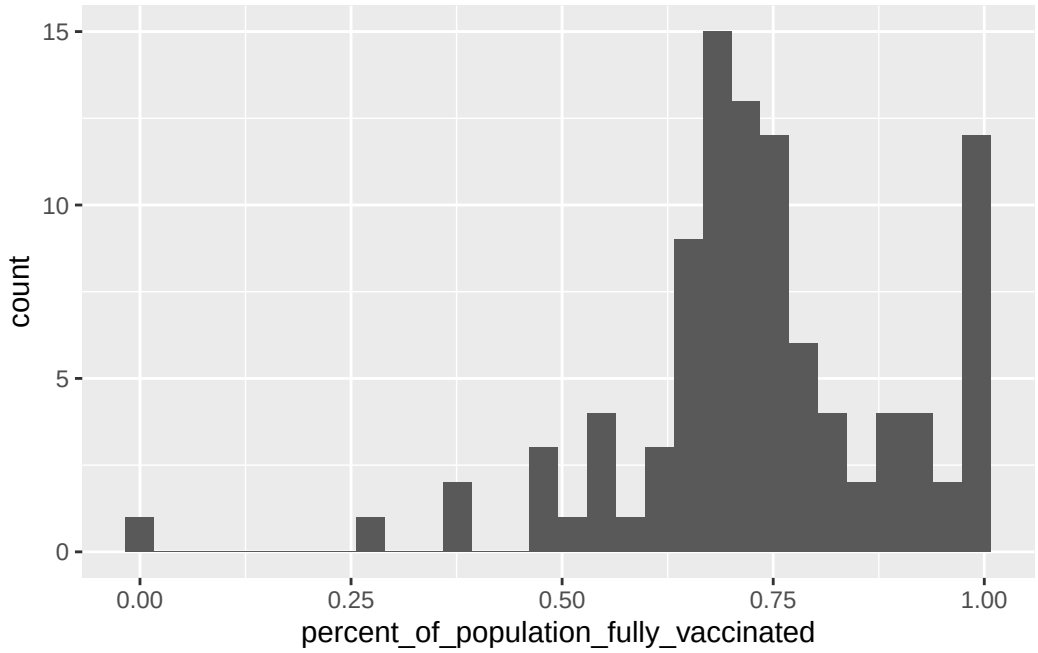
```
73.69 %
```

note: Barry's value will be different b/c he has different dataset but it should still be close

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"?

```
library(ggplot2)
ggplot(data=sd.2022.11.15) + aes(x=percent_of_population_fully_vaccinated) + geom_histogra
```

```
Warning: Removed 8 rows containing non-finite values (`stat_bin()`).
```



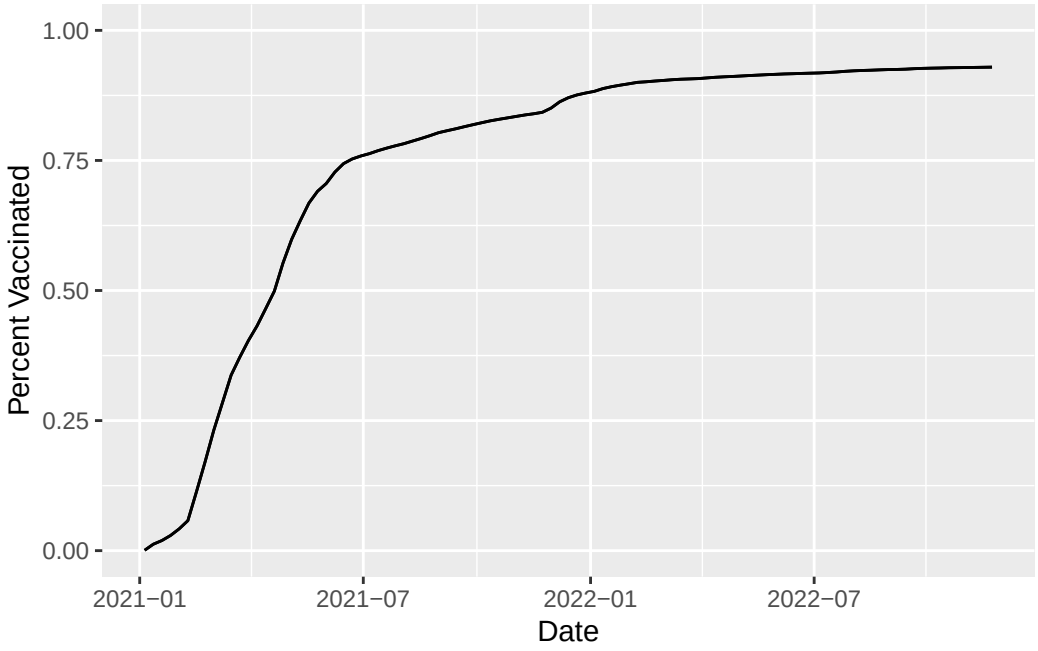
Focus on UCSD/La Jolla

use ucsd zip code to filter for UCSD/la jolla zip code

```
ucsd <- filter(sd, zip_code_tabulation_area == "92037")
```

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

```
ggplot(ucsd) +
  aes(x=as_of_date,
      y=percent_of_population_fully_vaccinated) +
  geom_line() +
  geom_line(group=1) +
  ylim(c(0,1)) +
  labs(x="Date", y="Percent Vaccinated")
```



Comparing to similar sized areas

```
#filter vax to 92037 zip and 2022-02-22 date
population.92037.20220222 <- filter(vax,zip_code_tabulation_area == "92037" & as_of_date =
head(population.92037.20220222)
```

```
  as_of_date zip_code_tabulation_area local_health_jurisdiction    county
1 2022-02-22                   92037                San Diego San Diego
  vaccine_equity_metric_quartile               vem_source
```

```
1                             4 Healthy Places Index Score
  age12_plus_population age5_plus_population tot_population
1               33675.6                36144          38168
  persons_fully_vaccinated persons_partially_vaccinated
1                    34452                          4084
  percent_of_population_fully_vaccinated
1                               0.902641
  percent_of_population_partially_vaccinated
1                                 0.107001
  percent_of_population_with_1_plus_dose booster_recip_count
1                                      1               12993
  bivalent_dose_recip_count eligible_recipient_count redacted
1                        NA                    34451       No
```

```
#subset to CA areas as of date 2022-11-15 > population 92037 zip 2022-02-22 date
vax.36 <- filter(vax, age5_plus_population > population.92037.20220222$age5_plus_populatio
```

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

```
  as_of_date zip_code_tabulation_area local_health_jurisdiction        county
1 2022-11-15                   92236                 Riverside     Riverside
2 2022-11-15                   92130                 San Diego     San Diego
3 2022-11-15                   94121             San Francisco San Francisco
4 2022-11-15                   94551                   Alameda       Alameda
5 2022-11-15                   94112             San Francisco San Francisco
6 2022-11-15                   94303               Santa Clara   Santa Clara
  vaccine_equity_metric_quartile                vem_source
1                              1 Healthy Places Index Score
2                              4 Healthy Places Index Score
3                              4 Healthy Places Index Score
4                              4 Healthy Places Index Score
5                              3 Healthy Places Index Score
6                              3 Healthy Places Index Score
  age12_plus_population age5_plus_population tot_population
1               38505.3                42923          45477
2               46300.3                53102          56134
3               39105.0                41363          43616
4               38947.9                43399          47227
5               75681.8                81107          84707
6               40033.3                44989          48244
  persons_fully_vaccinated persons_partially_vaccinated
```

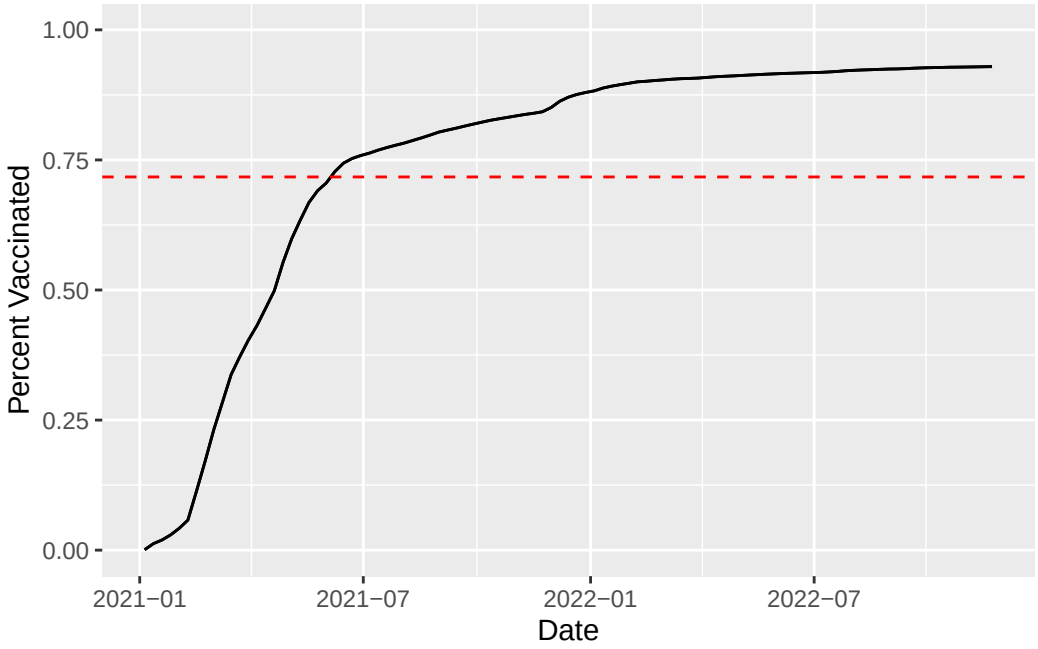
```
1                    30465                          3858
2                    52380                          5751
3                    36566                          2373
4                    32557                          2333
5                    78358                          4646
6                    41275                          4175
  percent_of_population_fully_vaccinated
1                               0.669899
2                               0.933124
3                               0.838362
4                               0.689373
5                               0.925048
6                               0.855547
  percent_of_population_partially_vaccinated
1                                   0.084834
2                                   0.102451
3                                   0.054407
4                                   0.049400
5                                   0.054848
6                                   0.086539
  percent_of_population_with_1_plus_dose booster_recip_count
1                               0.754733               12943
2                               1.000000               34821
3                               0.892769               28345
4                               0.738773               20223
5                               0.979896               56744
6                               0.942086               26288
  bivalent_dose_recip_count eligible_recipient_count redacted
1                      1395                    30375       No
2                     11203                    51780       No
3                     10994                    36013       No
4                      5568                    32234       No
5                     16019                    77580       No
6                      8573                    40853       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 "2022-11-15". Add this as a straight horizontal line to your plot from above with the geom_hline() function?

```
perc.pop.vax.greater92037 <- mean(vax.36$percent_of_population_fully_vaccinated,na.rm = TR
perc.pop.vax.greater92037
```

```
[1] 0.7172851
```

```
ggplot(ucsd) +
  aes(x=as_of_date,
      y=percent_of_population_fully_vaccinated) +
  geom_line() +
  geom_line(group=1) +
  ylim(c(0,1)) +
  labs(x="Date", y="Percent Vaccinated") +
  geom_hline(yintercept = perc.pop.vax.greater92037,col="red",linetype="dashed")
```



Note: value slightly different than Barry's because different dataset. values are close though

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)
```

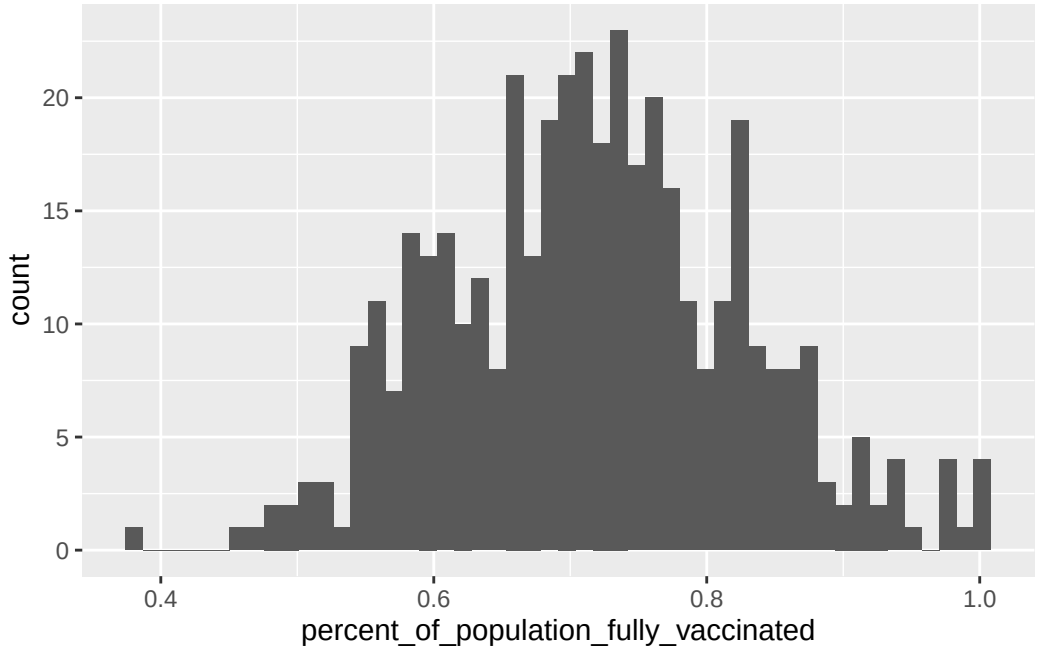
```
  as_of_date         zip_code_tabulation_area local_health_jurisdiction
Min.   :2022-11-15   Min.   :90001            Length:411
```

```
1st Qu.:2022-11-15   1st Qu.:91762            Class :character
Median :2022-11-15   Median :92646            Mode  :character
Mean   :2022-11-15   Mean   :92862
3rd Qu.:2022-11-15   3rd Qu.:94517
Max.   :2022-11-15   Max.   :96003
   county          vaccine_equity_metric_quartile  vem_source
Length:411         Min.   :1.000                   Length:411
Class :character   1st Qu.:1.000                   Class :character
Mode  :character   Median :2.000                   Mode  :character
                   Mean   :2.353
                   3rd Qu.:3.000
                   Max.   :4.000
age12_plus_population age5_plus_population tot_population
Min.   :31651         Min.   : 36181       Min.   : 38007
1st Qu.:37694         1st Qu.: 41612       1st Qu.: 44393
Median :43985         Median : 48573       Median : 52212
Mean   :46847         Mean   : 52012       Mean   : 55641
3rd Qu.:53932         3rd Qu.: 59168       3rd Qu.: 62910
Max.   :88557         Max.   :101902       Max.   :111165
persons_fully_vaccinated persons_partially_vaccinated
Min.   :17422            Min.   : 1733
1st Qu.:31926            1st Qu.: 2813
Median :37064            Median : 3542
Mean   :39837            Mean   : 4078
3rd Qu.:45034            3rd Qu.: 4666
Max.   :87151            Max.   :39160
percent_of_population_fully_vaccinated
Min.   :0.3785
1st Qu.:0.6396
Median :0.7155
Mean   :0.7173
3rd Qu.:0.7880
Max.   :1.0000
percent_of_population_partially_vaccinated
Min.   :0.04153
1st Qu.:0.05713
Median :0.06466
Mean   :0.07342
3rd Qu.:0.07717
Max.   :0.97744
percent_of_population_with_1_plus_dose booster_recip_count
Min.   :0.4390                         Min.   : 8603
1st Qu.:0.7083                         1st Qu.:17134
```

```
Median :0.7850                          Median :21640
Mean   :0.7851                          Mean   :22817
3rd Qu.:0.8594                          3rd Qu.:27266
Max.   :1.0000                          Max.   :56744
bivalent_dose_recip_count eligible_recipient_count   redacted
Min.   : 1375             Min.   :17321            Length:411
1st Qu.: 3418             1st Qu.:31820            Class :character
Median : 4941             Median :36758            Mode  :character
Mean   : 5619             Mean   :39609
3rd Qu.: 7270             3rd Qu.:44904
Max.   :16829             Max.   :86696
```

Q18. Using ggplot generate a histogram of this data.

```
ggplot(data=vax.36) + aes(x=percent_of_population_fully_vaccinated) + geom_histogram(bins=
```



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

```
vax.9204<-vax %>% filter(as_of_date == "2022-11-15") %>%
  filter(zip_code_tabulation_area=="92040") %>%
```

```
  select(percent_of_population_fully_vaccinated)
if (vax.9204<0.7172851){
  print("92040 is Below")} else {
    print("92040 is Above")
  }
```

```
[1] "92040 is Below"
```

```
vax.92109<-vax %>% filter(as_of_date == "2022-11-15") %>%
  filter(zip_code_tabulation_area=="92109") %>%
  select(percent_of_population_fully_vaccinated)
if (vax.92109<0.7172851){
  print("92109 is Below")} else {
    print("92109 is Above")
  }
```

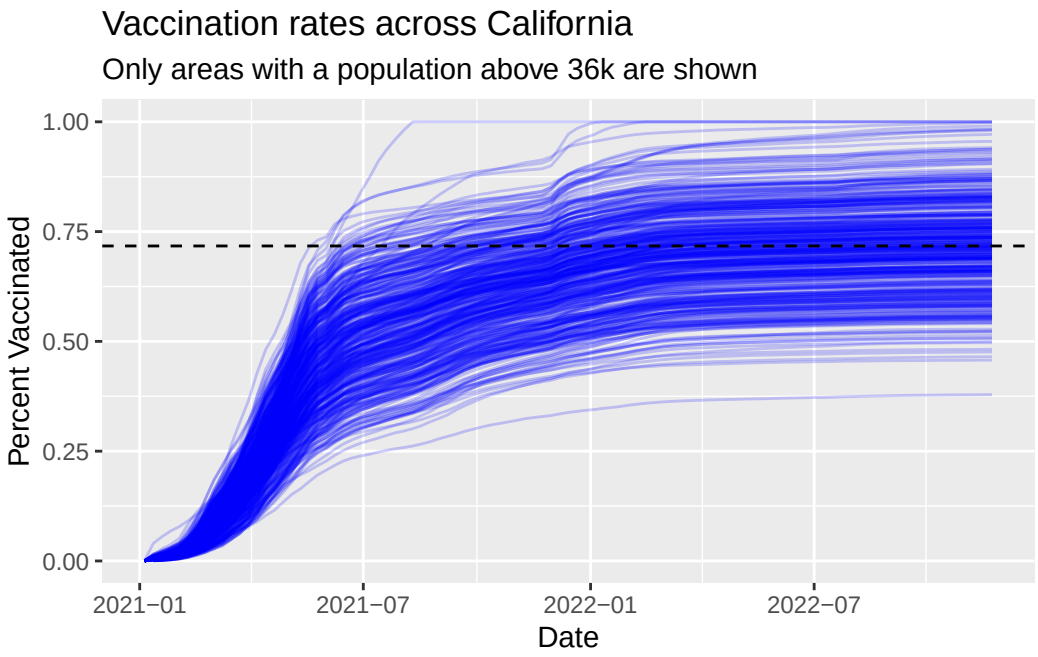
```
[1] "92109 is Below"
```

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

```
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 rates across California",
       subtitle="Only areas with a population above 36k are shown") +
  geom_hline(yintercept = perc.pop.vax.greater92037, linetype="dashed")
```

```
Warning: Removed 184 rows containing missing values (`geom_line()`).
```



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

Not horrible, but definitely not great lol. There are a ton of places with lower than 75% vaccinated. It would be interesting to probe further and see if regions could be grouped together to show larger areas that are very good about being vaccinated vs. very bad.

```
#report session info
sessionInfo()
```

```
R version 4.2.1 (2022-06-23)
Platform: x86_64-apple-darwin17.0 (64-bit)
Running under: macOS Big Sur ... 10.16

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.2/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_3.4.0    dplyr_1.0.10     zipcodeR_0.3.5   lubridate_1.9.0
[5] timechange_0.1.1

loaded via a namespace (and not attached):
 [1] httr_1.4.4         tidyr_1.2.1        bit64_4.0.5        jsonlite_1.8.3
 [5] assertthat_0.2.1   sp_1.5-1           highr_0.9          blob_1.2.3
 [9] yaml_2.3.6         tidycensus_1.2.3   pillar_1.8.1       RSQLite_2.2.18
[13] lattice_0.20-45    glue_1.6.2         uuid_1.1-0         digest_0.6.30
[17] rvest_1.0.3        colorspace_2.0-3   htmltools_0.5.3    pkgconfig_2.0.3
[21] raster_3.6-3       purrr_0.3.5        scales_1.2.1       terra_1.6-41
[25] tzdb_0.3.0         tigris_1.6.1       tibble_3.1.8       proxy_0.4-27
[29] farver_2.1.1       generics_0.1.3     ellipsis_0.3.2     cachem_1.0.6
[33] withr_2.5.0        repr_1.1.4         skimr_2.1.4        cli_3.4.1
[37] magrittr_2.0.3     crayon_1.5.2       memoise_2.0.1      maptools_1.1-5
[41] evaluate_0.18      fansi_1.0.3        xml2_1.3.3         foreign_0.8-83
[45] class_7.3-20       tools_4.2.1        hms_1.1.2          lifecycle_1.0.3
[49] stringr_1.4.1      munsell_0.5.0      compiler_4.2.1     e1071_1.7-12
[53] rlang_1.0.6        classInt_0.4-8     units_0.8-0        grid_4.2.1
[57] rstudioapi_0.14    rappdirs_0.3.3     labeling_0.4.2     base64enc_0.1-3
[61] rmarkdown_2.18     gtable_0.3.1       codetools_0.2-18   DBI_1.1.3
[65] curl_4.3.3         R6_2.5.1           knitr_1.40         rgdal_1.6-2
[69] fastmap_1.1.0      bit_4.0.4          utf8_1.2.2         KernSmooth_2.23-20
[73] readr_2.1.3        stringi_1.7.8      Rcpp_1.0.9         vctrs_0.5.0
[77] sf_1.0-9           tidyselect_1.2.0   xfun_0.34
```