Project14

Haroon Riyaz (PID A15377799)

3/7/2022

Read vaccination data

```
vax <- read.csv("vacdata.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”

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

Q4. What is the latest date in this dataset?

2022-03-01

Use the skim() function to get overview of data

```r
library(skimr)
numcol_skim <- skimr::skim(vax)
numcol_skim
```

Table: 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
zip_code_tabulation_area	0	1.00	93665.11	1817.39	90001	9225
vaccine_equity_metric_quartile	5307	0.95	2.44	1.11	1
age12_plus_population	0	1.00	18895.04	18993.91	0	134
age5_plus_population	0	1.00	20875.24	21106.02	0	146
persons_fully_vaccinated	18338	0.83	12155.61	13063.88	11	106
persons_partially_vaccinated	18338	0.83	831.74	1348.68	11	7
percent_of_population_fully_vaccinated	18338	0.83	0.51	0.26	0
percent_of_population_partially_vaccinated	18338	0.83	0.05	0.09	0
percent_of_population_with_1_plus_dose	18338	0.83	0.54	0.28	0
booster_recip_count	64317	0.40	4100.55	5900.21	11	17

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

There are *9* numeric columns (excludes skim_variable and hist)

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

```r
NA_PFV <- sum(is.na(vax$persons_fully_vaccinated))
NA_PFV
```

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

18338 with current values (*18174* on lab sheet)

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

```r
# Divide # of missing values by # of rows, then round
round((NA_PFV/nrow(vax))*100, 2)
```

```
## [1] 17.04
```

```r
# Lab Sheet Values
round((18174/105840)*100, 2)
```

```
## [1] 17.17
```

17.04% with current values (*17.17%* on lab sheet)

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

The data is missing since it represents people who haven't gotten the vaccine yet.

Using lubridate to Simplify Dates

```r
library(lubridate)
```

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

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

See Today's Date

```r
today()
```

```
## [1] "2022-03-07"
```

```r
# Specify Date Format (ymd)

vax$as_of_date <- ymd(vax$as_of_date)
```

Now math can be done with the dates!

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

```
## Time difference of 426 days
```

How many days the dataset spans

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

```
## Time difference of 420 days
```

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

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

```
## Time difference of 6 days
```

6 days

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

```
table(vax$as_of_date)
```

```
##
## 2021-01-05 2021-01-12 2021-01-19 2021-01-26 2021-02-02 2021-02-09 2021-02-16
##       1764       1764       1764       1764       1764       1764       1764
## 2021-02-23 2021-03-02 2021-03-09 2021-03-16 2021-03-23 2021-03-30 2021-04-06
##       1764       1764       1764       1764       1764       1764       1764
## 2021-04-13 2021-04-20 2021-04-27 2021-05-04 2021-05-11 2021-05-18 2021-05-25
##       1764       1764       1764       1764       1764       1764       1764
## 2021-06-01 2021-06-08 2021-06-15 2021-06-22 2021-06-29 2021-07-06 2021-07-13
##       1764       1764       1764       1764       1764       1764       1764
## 2021-07-20 2021-07-27 2021-08-03 2021-08-10 2021-08-17 2021-08-24 2021-08-31
##       1764       1764       1764       1764       1764       1764       1764
## 2021-09-07 2021-09-14 2021-09-21 2021-09-28 2021-10-05 2021-10-12 2021-10-19
##       1764       1764       1764       1764       1764       1764       1764
## 2021-10-26 2021-11-02 2021-11-09 2021-11-16 2021-11-23 2021-11-30 2021-12-07
##       1764       1764       1764       1764       1764       1764       1764
## 2021-12-14 2021-12-21 2021-12-28 2022-01-04 2022-01-11 2022-01-18 2022-01-25
##       1764       1764       1764       1764       1764       1764       1764
## 2022-02-01 2022-02-08 2022-02-15 2022-02-22 2022-03-01
##       1764       1764       1764       1764       1764
```

```
nrow(table(vax$as_of_date))
```

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

61 unique dates (lab sheet value different)

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.
```

Calculate distance between 2 points with zip codes

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

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

Review Census Data

```
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 San Diego Area

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

Using 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] 6527
```

dplyr useful for subsetting

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

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

```
filtered_SD <- filter(vax, county == "San Diego")
length(unique(filtered_SD$zip_code_tabulation_area))
```

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

107 distinct zip codes in SD

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

```
maxSD12 <- which.max(filtered_SD$age12_plus_population)
filtered_SD$zip_code_tabulation_area[maxSD12]
```

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

92154 is the zip code with largest 12+ pop.

Q13. What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2022-02-22"?

```
PPFV_SD_02.22 <- filter(vax, county == "San Diego", as_of_date == "2022-02-22")

mean(PPFV_SD_02.22$percent_of_population_fully_vaccinated, na.rm = TRUE)
```

```
## [1] 0.7041551
```

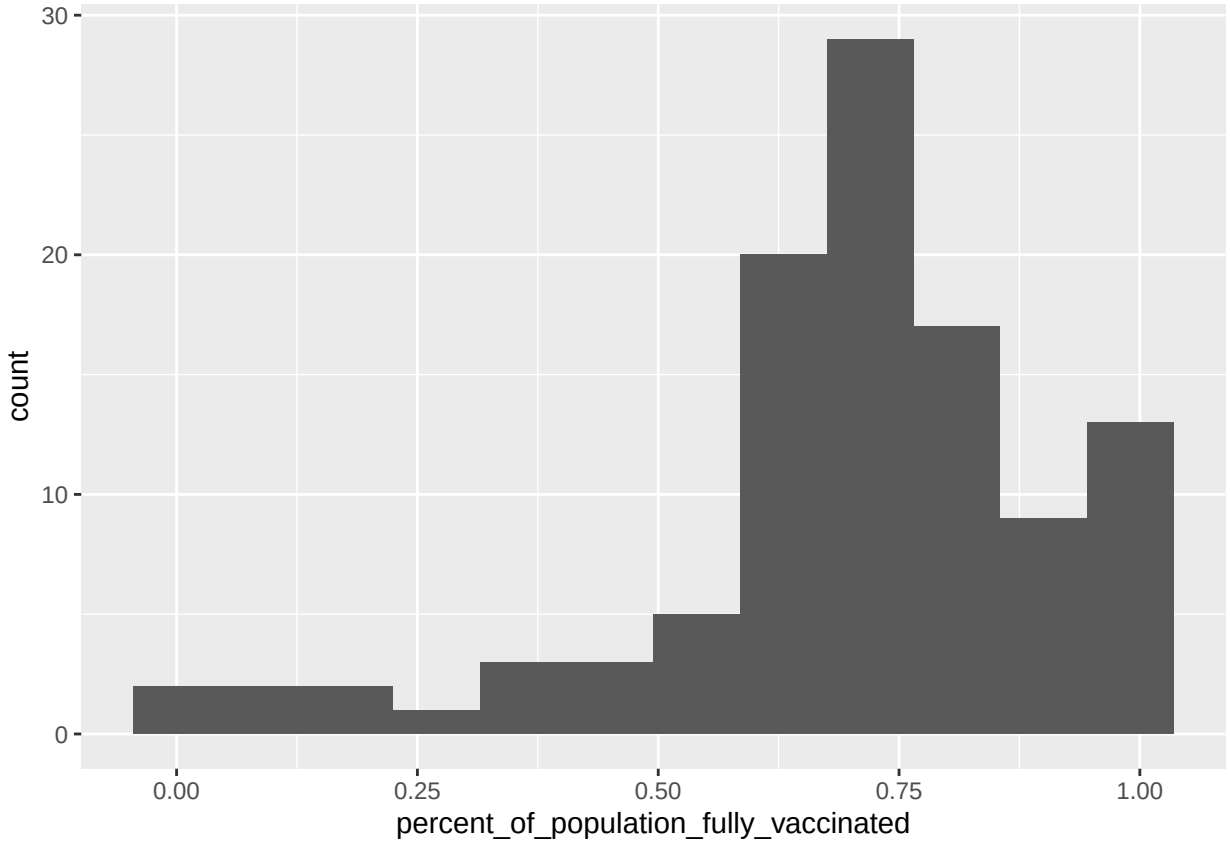
70.42%

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"?

```
library(ggplot2)

ggplot(PPFV_SD_02.22, aes(percent_of_population_fully_vaccinated)) +
geom_histogram(bins = 12)
```

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



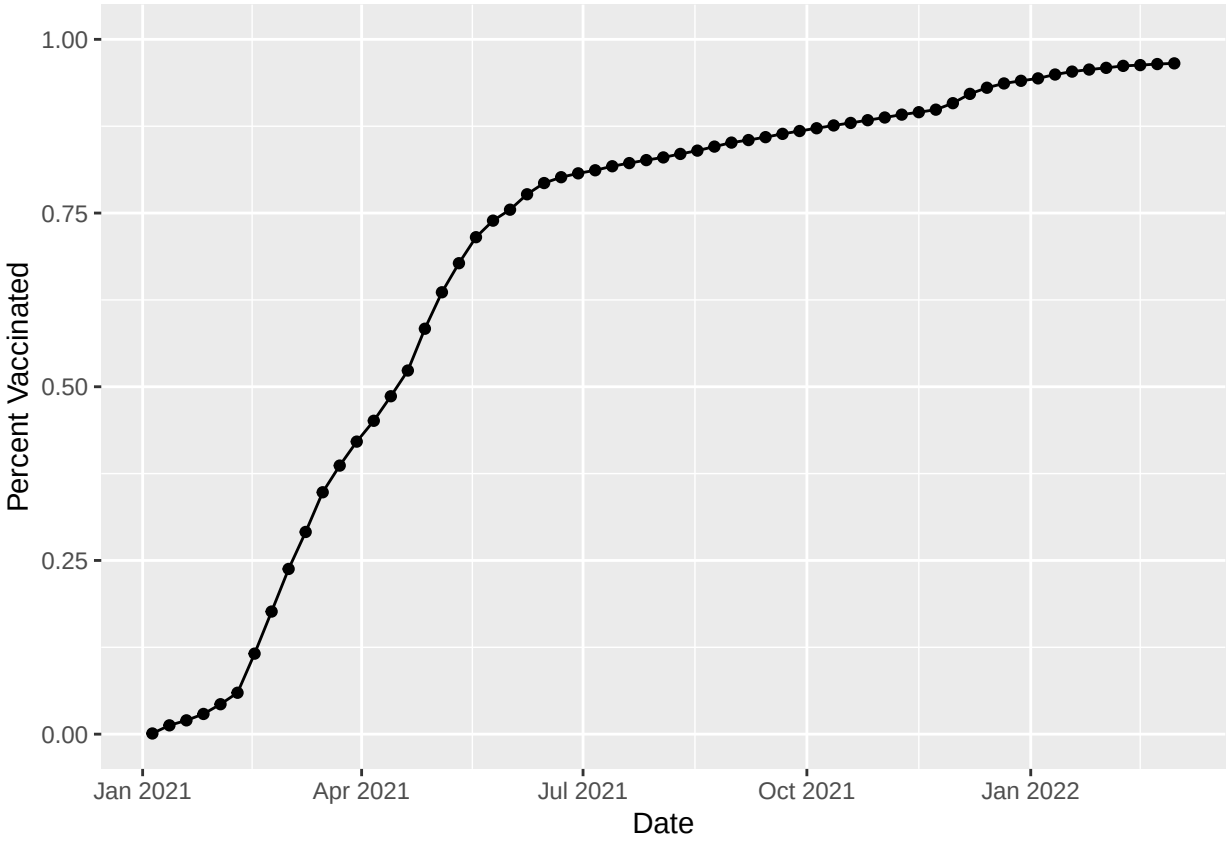
Focus on UCSD

```
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:

```
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")
```



Comparing to similar sized areas

```
# CA areas with population as large as 92037
vax.36 <- filter(vax, age5_plus_population > 36144 &
                 as_of_date == "2022-02-22")

# 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 "2022-02-22". Add this as a straight horizontal line to your plot from above with the geom_hline() function?

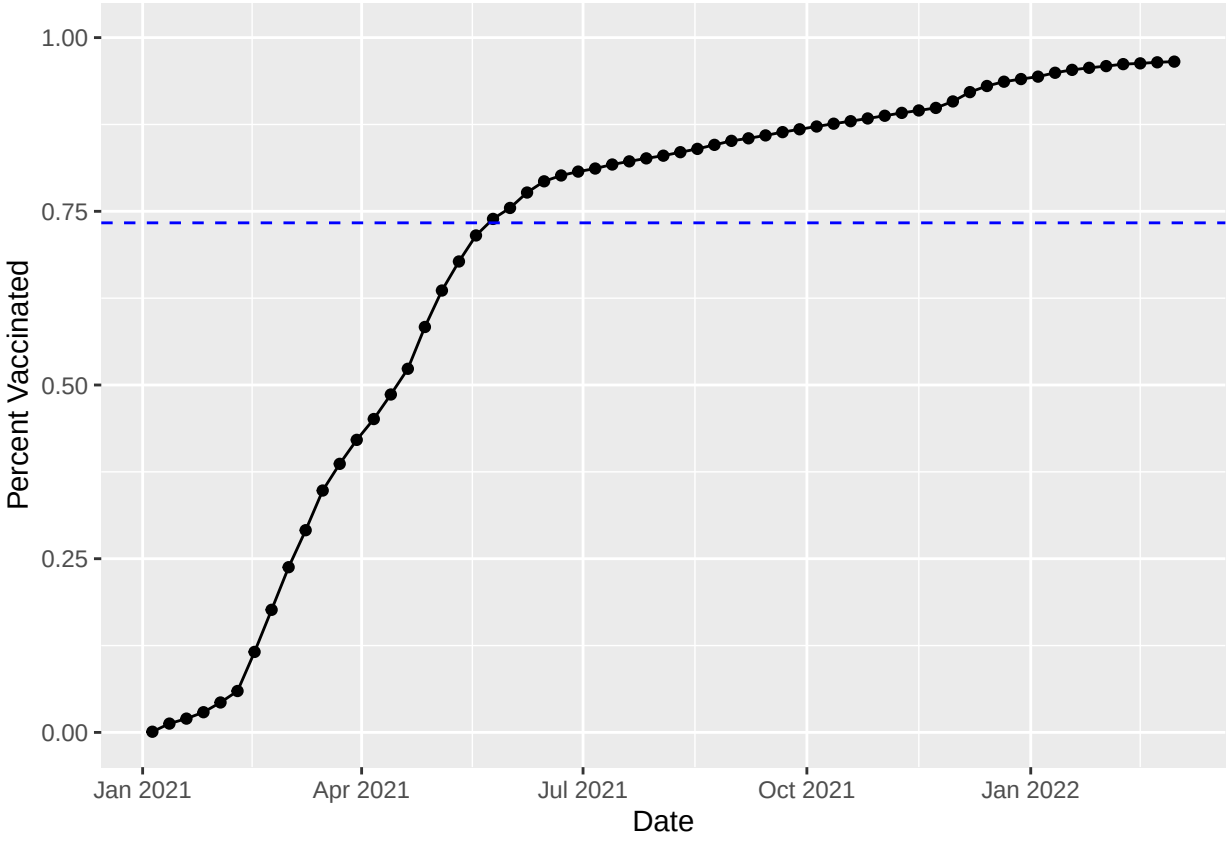
```
mean_vax.36 <- mean(vax.36$percent_of_population_fully_vaccinated, na.rm = TRUE)
mean_vax.36
```

```
## [1] 0.733385
```

73.34% is the mean percentage of population fully vaccinated for zip codes with pop. as large as 92037

```
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") +
  geom_hline(yintercept = mean_vax.36, linetype = "dashed", color = "blue")
```



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 "2022-02-22"?

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

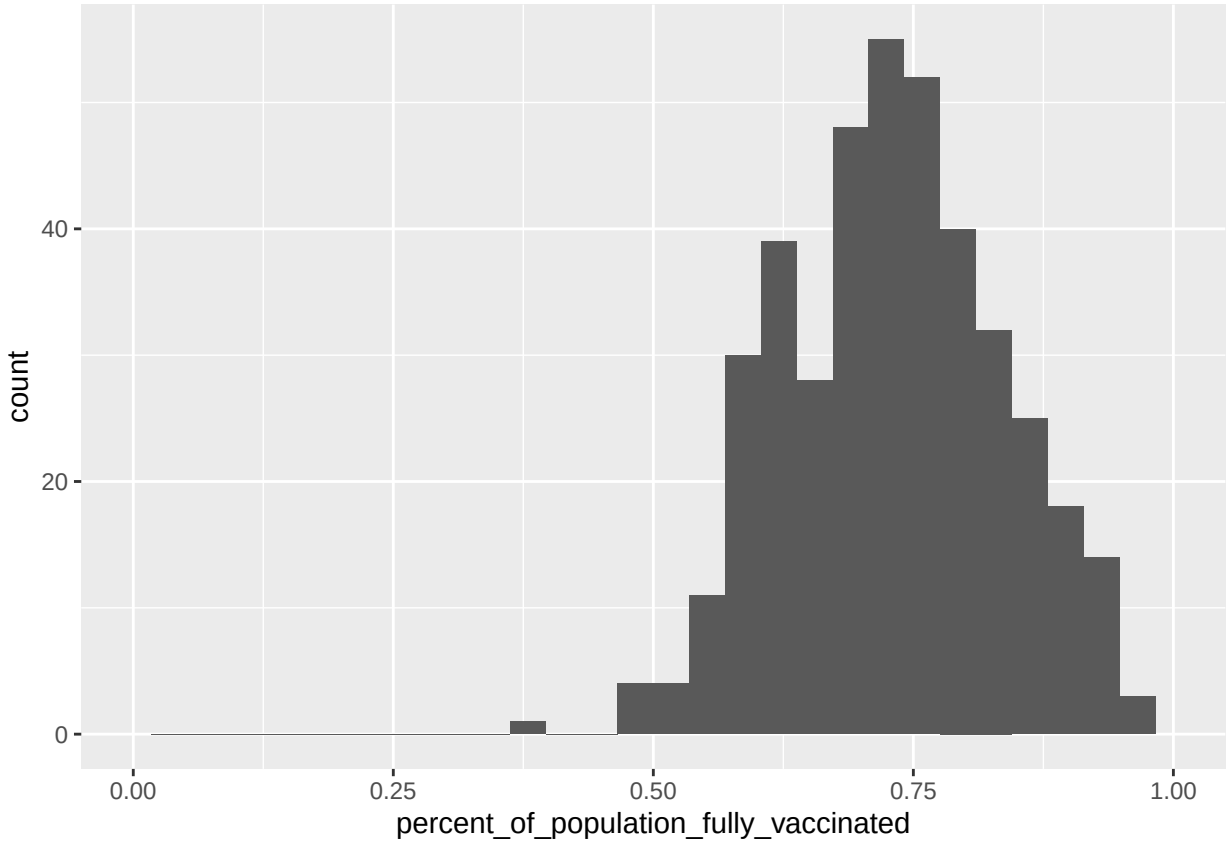
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3881  0.6539  0.7333  0.7334  0.8027  1.0000
```

Q18. Using ggplot generate a histogram of this data.

```
ggplot(vax.36, aes(percent_of_population_fully_vaccinated)) +
geom_histogram() + xlim(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?

```
vax %>% filter(as_of_date == "2022-02-22") %>%
  filter(zip_code_tabulation_area =="92109") %>%
  select(percent_of_population_fully_vaccinated)
```

```
##   percent_of_population_fully_vaccinated
## 1                               0.723044
```

```
vax %>% filter(as_of_date == "2022-02-22") %>%
  filter(zip_code_tabulation_area =="92040") %>%
  select(percent_of_population_fully_vaccinated)
```

```
##   percent_of_population_fully_vaccinated
## 1                               0.551304
```

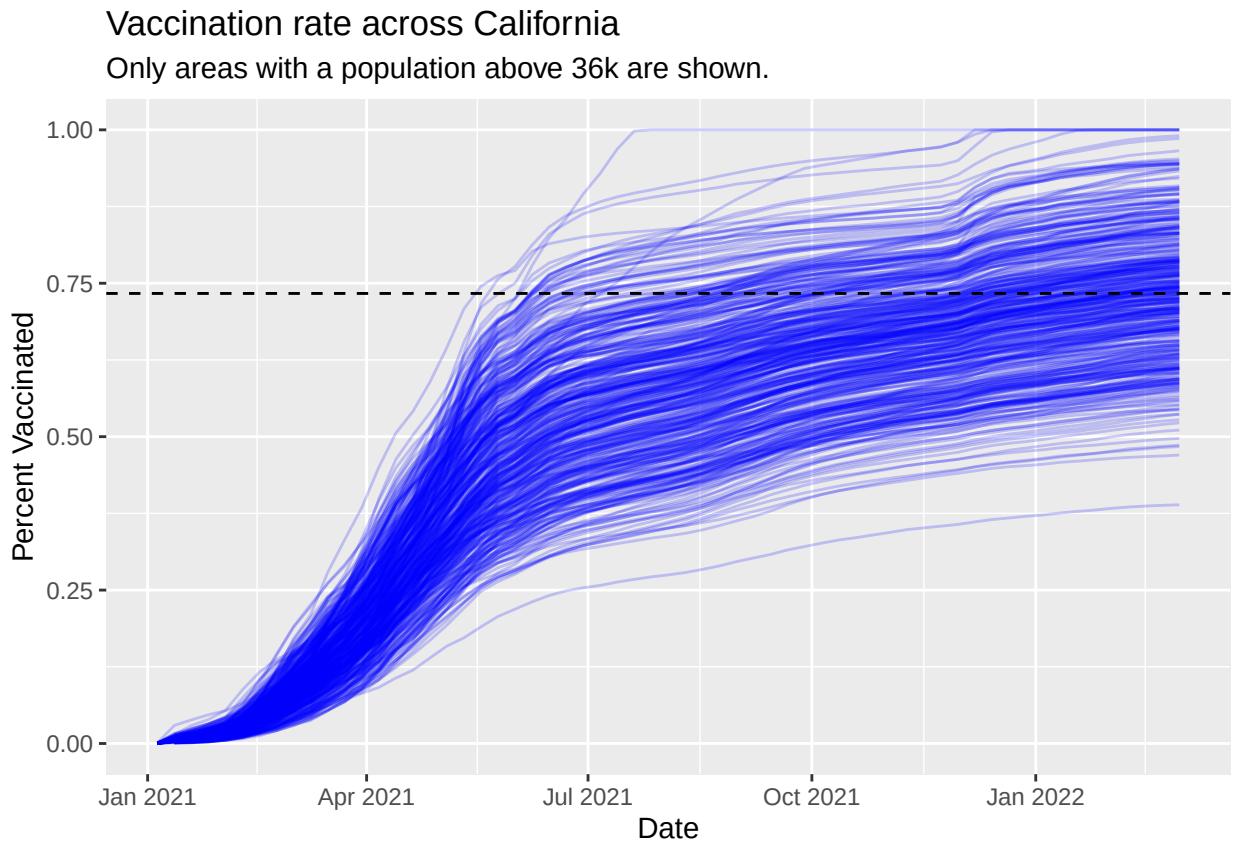
Both ZIP code areas (92109 and 92040) are below the average value (0.7230 and 0.5513 $<$ .7334)

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 = mean_vax.36, linetype = "dashed")
```

```
## 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?

I would feel great but this is my last quarter at UCSD so I won't be returning to in-person classes afterwards.