Class14_Mini Project Vaccination

Morgan Farrell

3/4/2022

Import dataset for San Diego County Vaccination Status

```r
# Import vaccination data
vax <- read.csv("covid19vaccinesbyzipcode.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?

3 persons_fully_vaccinated

Q2. What column details the Zip code tabulation area?

12 zip_code_tabulation_area

Find the dates

```r
#first date
vax$as_of_date[1]
```

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

```r
#latest date
vax$as_of_date[nrow(vax)]
```

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

```r
#install.packages("lubridate")
library(lubridate)
```

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

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

Q3. What is the earliest date in this dataset?

2021-01-05

Q4. What is the latest date in this dataset?

2022-03-01

Using the skimr package to see the dataset characteristics

```r
#install.packages("skimr")
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_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?

10, but Zip code is counted as numeric so it should really be 9

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

18338

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

```
(18338/sum(vax$persons_fully_vaccinated, na.rm=TRUE))*100
```

```
## [1] 0.00169001
```

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

Because they put kids vaccination in different columns?

Using the package 'lubridate()'

```
library(lubridate)
time_length(today()-ymd("1994-03-06"), "years") #I'm getting old
```

```
## [1] 27.99452
```

Store the as_of_date column as a variable so we can do math with it

```
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[1]
```

```
## Time difference of 423 days
```

Days between the first and last vaccination

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

```
## Time difference of 420 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] 61
```

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

Figuring out SD counties using base R

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

Using 'dplyr' package to filter the County data instead

```
library(tidyverse) #this has a bunch of packages ggplot, dplyr, etc.
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x lubridate::as.difftime() masks base::as.difftime()
## x lubridate::date()        masks base::date()
## x dplyr::filter()          masks stats::filter()
## x lubridate::intersect()   masks base::intersect()
## x dplyr::lag()             masks stats::lag()
## x lubridate::setdiff()     masks base::setdiff()
## x lubridate::union()       masks base::union()
```

Using the filter function

```
sd <- filter(vax, county== "San Diego")
```

Using the pipe funtion to pass the variable onto the arguments

```
#vax %>% filter(county == "San Diego")
```

```
head(sd, 3)
```

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 1 2021-01-05                    92130                 San Diego San Diego
## 2 2021-01-05                    91945                 San Diego San Diego
## 3 2021-01-05                    91917                 San Diego San Diego
##   vaccine_equity_metric_quartile                 vem_source
## 1                              4 Healthy Places Index Score
## 2                              2 Healthy Places Index Score
## 3                              1    CDPH-Derived ZCTA Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1               46300.3                53102                       61
## 2               22820.5                25486                       NA
## 3                 826.1                  939                       NA
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                           27                               0.001149
## 2                           NA                                     NA
## 3                           NA                                     NA
##   percent_of_population_partially_vaccinated
## 1                                   0.000508
## 2                                         NA
## 3                                         NA
##   percent_of_population_with_1_plus_dose booster_recip_count
## 1                               0.001657                  NA
## 2                                     NA                  NA
## 3                                     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
```

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

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

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

```
length(table(sd$zip_code_tabulation_area))
```

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

Using dplyr and the pipe function

```
vax %>%
  filter(county == "San Diego") %>%
  select(zip_code_tabulation_area) %>%
  unique() %>%
  nrow()
```

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

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

92154

```
#vax %>%
#  filter(county == "San Diego") %>%
#  select(age12_plus_population) %>%
#  order(decreasing = TRUE)
```

Base R way to answer the question

```
inds <- order(sd$age12_plus_population, decreasing = TRUE)
sd[inds[1],]
```

```
##    as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 91 2021-01-05                    92154                 San Diego San Diego
##    vaccine_equity_metric_quartile                 vem_source
## 91                              2 Healthy Places Index Score
##    age12_plus_population age5_plus_population persons_fully_vaccinated
## 91               76365.2                82971                       18
##    persons_partially_vaccinated percent_of_population_fully_vaccinated
## 91                         22                             0.000217
##    percent_of_population_partially_vaccinated
## 91                                 0.000265
##    percent_of_population_with_1_plus_dose booster_recip_count
## 91                            0.000482                  NA
##                                                                   redacted
## 91 Information redacted in accordance with CA state privacy requirements
```

Using dplyr and the ‘arrange()’ function

```
head(arrange(sd, -age12_plus_population)) # the minus means sort opposite, which the default is lowest
```

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 1 2021-01-05                   92154                 San Diego San Diego
## 2 2021-01-12                   92154                 San Diego San Diego
## 3 2021-01-19                   92154                 San Diego San Diego
## 4 2021-01-26                   92154                 San Diego San Diego
## 5 2021-02-02                   92154                 San Diego San Diego
## 6 2021-02-09                   92154                 San Diego San Diego
##   vaccine_equity_metric_quartile                 vem_source
## 1                              2 Healthy Places Index Score
## 2                              2 Healthy Places Index Score
## 3                              2 Healthy Places Index Score
## 4                              2 Healthy Places Index Score
## 5                              2 Healthy Places Index Score
## 6                              2 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1               76365.2                82971                       18
## 2               76365.2                82971                      282
## 3               76365.2                82971                      671
## 4               76365.2                82971                      986
## 5               76365.2                82971                     1381
## 6               76365.2                82971                     2136
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                           22                               0.000217
## 2                           37                               0.003399
## 3                           93                               0.008087
## 4                          216                               0.011884
## 5                          432                               0.016644
## 6                          761                               0.025744
##   percent_of_population_partially_vaccinated
## 1                                   0.000265
## 2                                   0.000446
## 3                                   0.001121
## 4                                   0.002603
## 5                                   0.005207
## 6                                   0.009172
##   percent_of_population_with_1_plus_dose booster_recip_count
## 1                               0.000482                  NA
## 2                               0.003845                  NA
## 3                               0.009208                  NA
## 4                              0.014487                 NA
## 5                              0.021851                 NA
## 6                              0.034916                 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
```

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

70.53%

```
#Using the pipe method, a bit complicated
sd %>%
filter(as_of_date == "2022-03-01") %>%
select(percent_of_population_fully_vaccinated) %>%
colMeans(na.rm = T)
```

```
## percent_of_population_fully_vaccinated
##                              0.7052904
```

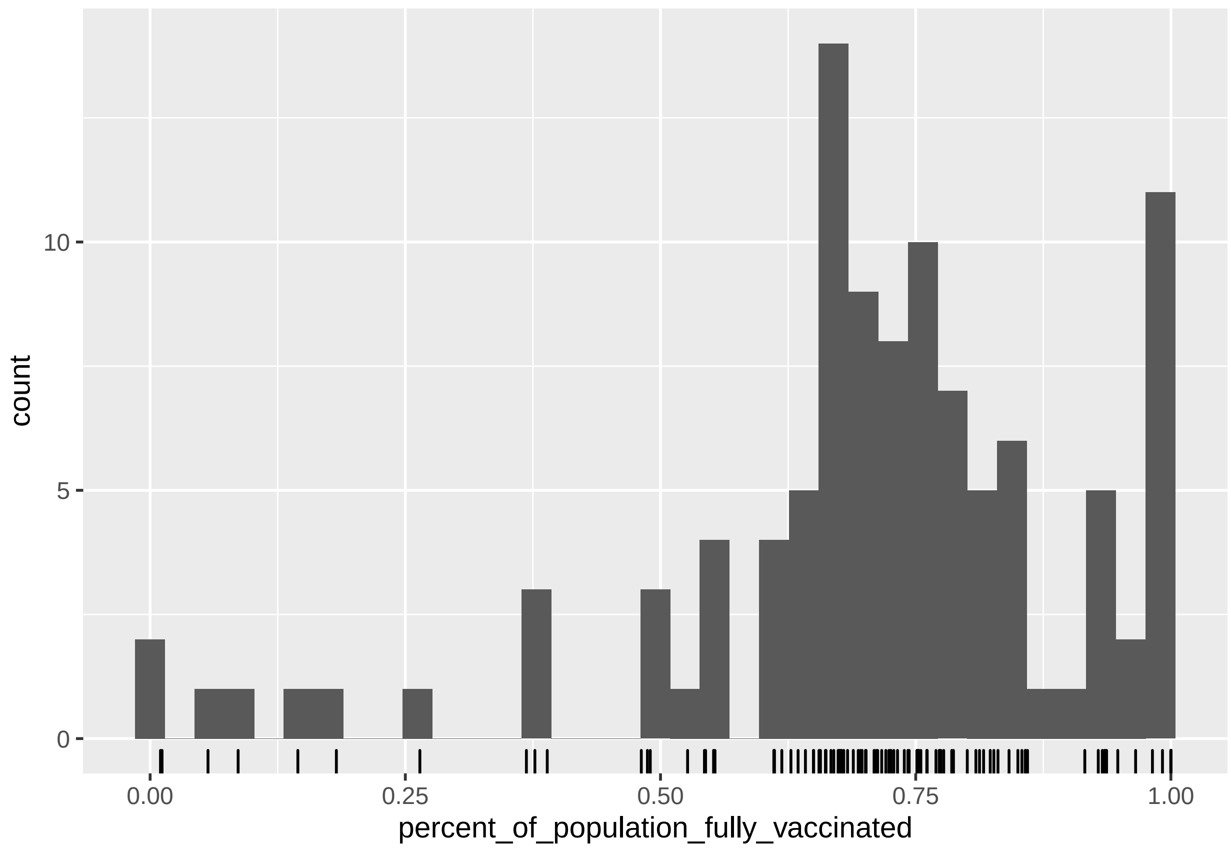
```
sd.now <- filter(sd, as_of_date == "2022-03-01")
sd.mean <- mean(sd.now$percent_of_population_fully_vaccinated, na.rm = TRUE)
sd.mean
```

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

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"?

```
ggplot(sd.now, aes(percent_of_population_fully_vaccinated))+
  geom_histogram(bins = 35) +
  geom_rug()
```

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



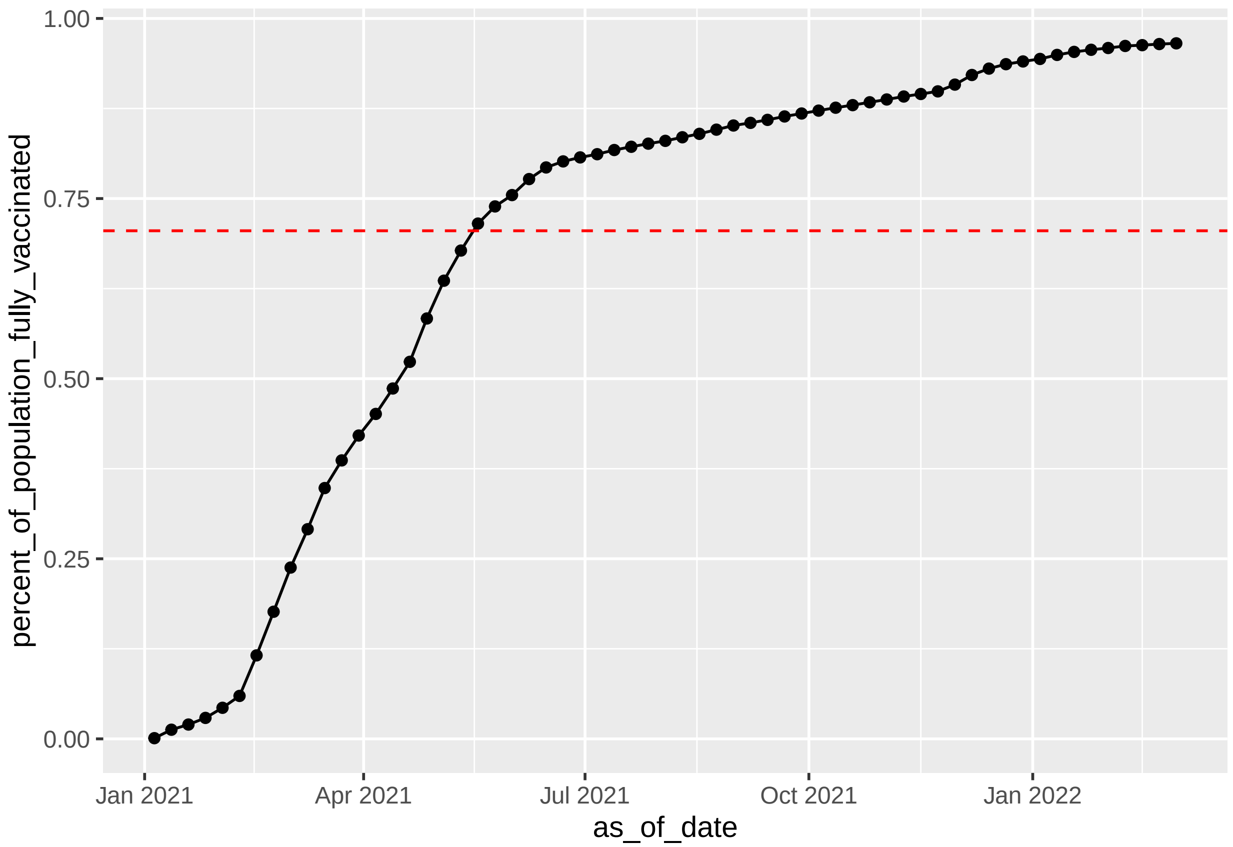
Compare UCSD to SD

```
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(x= as_of_date, y= percent_of_population_fully_vaccinated)) +
  geom_point() +
  geom_line(group=1) +
  geom_hline(yintercept= sd.mean, col="red", linetype=2)
```



```
  labs(x="Date", y="Percent Vaccinated", title = "Vaccination at UCSD")
```

```
## $x
## [1] "Date"
##
## $y
## [1] "Percent Vaccinated"
##
## $title
## [1] "Vaccination at UCSD"
##
## attr(,"class")
## [1] "labels"
```

Subset to all CA areas with a population as large as 92037

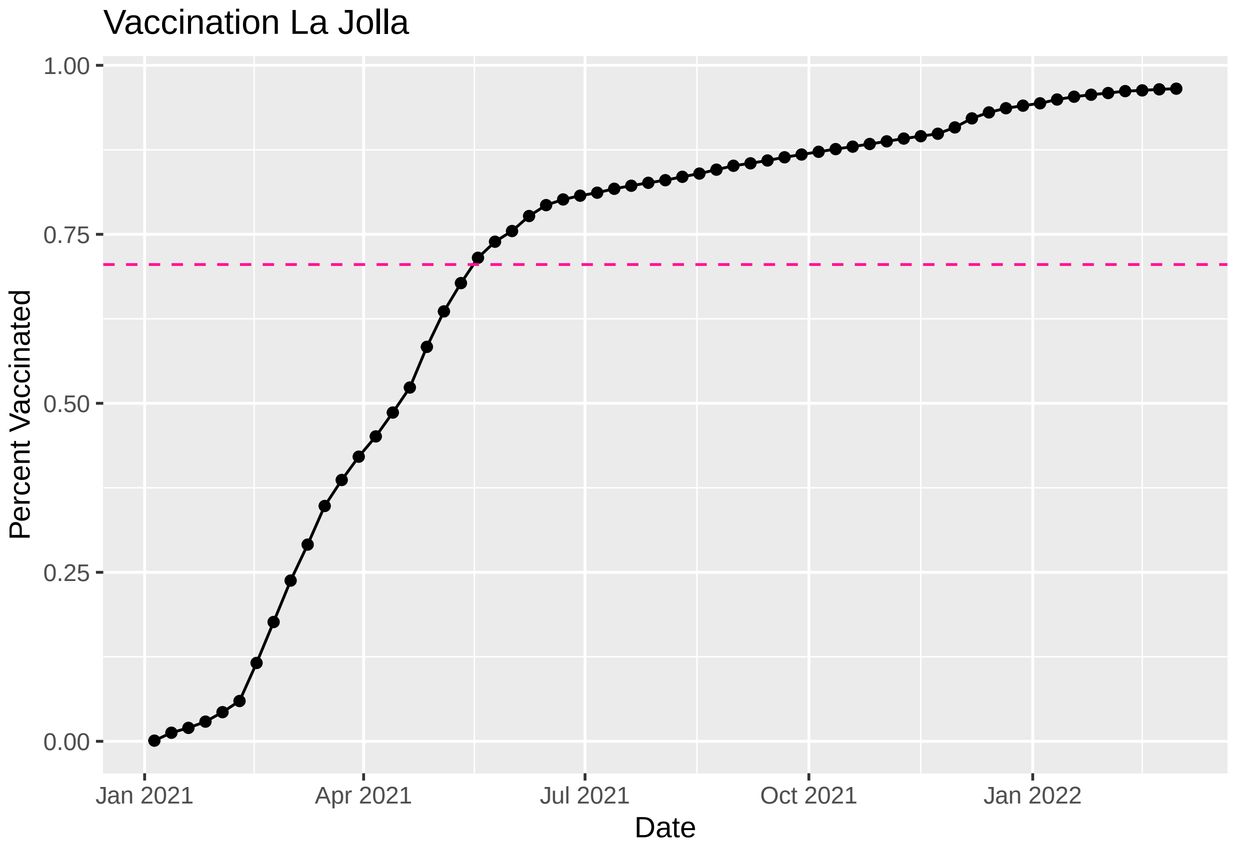
```
# Subset to all CA areas with a population as large as 92037
vax.36 <- filter(vax, age5_plus_population > 36144 &
                as_of_date == "2022-03-01")
```

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?

```
sd.36 <- filter(vax.36, as_of_date == "2022-03-01")
sd.36.mean <- mean(sd.now$percent_of_population_fully_vaccinated, na.rm = TRUE)
sd.36.mean
```

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

```
ggplot(ucsd, aes(x= as_of_date, y= percent_of_population_fully_vaccinated)) +
  geom_point() +
  geom_line(group=1) +
  geom_hline(yintercept= sd.36.mean, col="deep pink", linetype=2) +
  labs(x="Date", y="Percent Vaccinated", title = "Vaccination La Jolla")
```



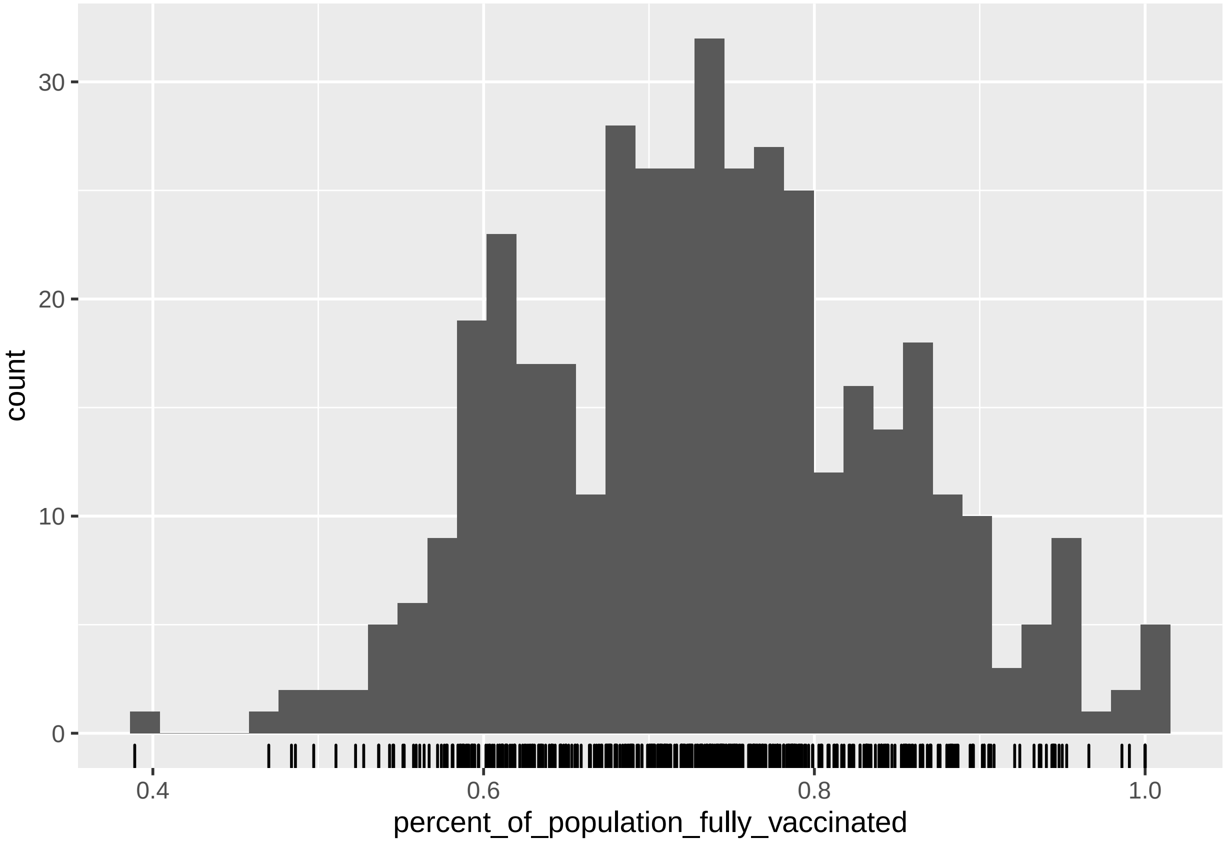
Q17. What is the 6 number summary (Min, 1st Qu., Median, Mean, 3rd Qu., and Max) of the “Percent of Population Fully Vaccinated” values for ZIP code areas with a population as large as 92037 (La Jolla) as_of_date “2022-02-22”?

```
summary.sd.36 <- summary(sd.36$percent_of_population_fully_vaccinated)
summary.sd.36
```

```
##    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(sd.36, aes(percent_of_population_fully_vaccinated))+
  geom_histogram(bins = 35) +
  geom_rug()
```



# 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=="92040") %>%
  select(percent_of_population_fully_vaccinated)
```

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

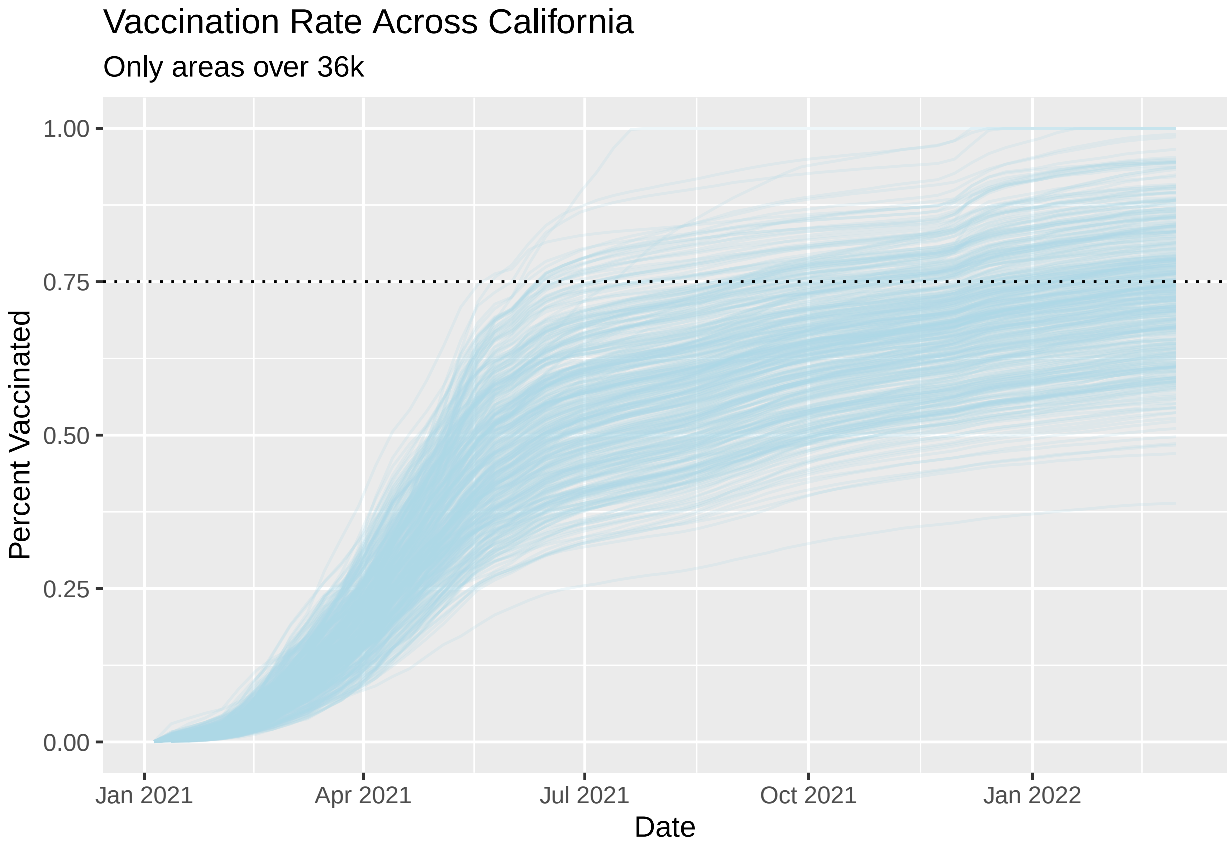
Q20. Finally make a time course plot of vaccination progress for all areas in the full dataset with a age5_plus_population > 36144.

```
vax.36.all <- filter(vax, age5_plus_population > 36144)
head(vax.36.all)
```

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction         county
## 1 2021-01-05                    92130                 San Diego      San Diego
## 2 2021-01-05                    91739            San Bernardino San Bernardino
## 3 2021-01-05                    91763            San Bernardino San Bernardino
## 4 2021-01-05                    92236                 Riverside      Riverside
## 5 2021-01-05                    94080                 San Mateo      San Mateo
## 6 2021-01-05                    94578                   Alameda        Alameda
##   vaccine_equity_metric_quartile                 vem_source
## 1                              4 Healthy Places Index Score
## 2                              4 Healthy Places Index Score
## 3                              1 Healthy Places Index Score
## 4                              1 Healthy Places Index Score
## 5                              4 Healthy Places Index Score
## 6                              2 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1               46300.3                53102                       61
## 2               33163.9                37166                       15
## 3               32730.4                36625                       NA
## 4               38505.3                42923                       NA
## 5               59769.6                64444                       NA
## 6               35092.5                38875                       NA
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                           27                               0.001149
## 2                           11                               0.000404
## 3                           NA                                     NA
## 4                           NA                                     NA
## 5                           NA                                     NA
## 6                           NA                                     NA
##   percent_of_population_partially_vaccinated
## 1                                   0.000508
## 2                                   0.000296
## 3                                         NA
## 4                                         NA
## 5                                         NA
## 6                                         NA
##   percent_of_population_with_1_plus_dose booster_recip_count
## 1                               0.001657                  NA
## 2                               0.000700                 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
```

```
ggplot(vax.36.all) +
  aes(as_of_date,
      percent_of_population_fully_vaccinated,
      group=zip_code_tabulation_area) +
  geom_line(alpha=0.2, color= "light blue") +
  ylim(c(0,1)) +
  labs(x="Date", y="Percent Vaccinated",
       title= "Vaccination Rate Across California",
       subtitle="Only areas over 36k") +
  geom_hline(yintercept = 0.75, linetype= 3)
```

```
## 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?

What Spring Break? Probably best not to travel though.