class17_bimm143

Hope (PID: A15652616)

11/23/2021

```
vax <- read.csv("covid19vaccinesbyzipcode_test.csv")
head(vax)
```

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 1 2021-01-05                   92804                     Orange    Orange
## 2 2021-01-05                   92626                     Orange    Orange
## 3 2021-01-05                   92250                   Imperial  Imperial
## 4 2021-01-05                   92637                     Orange    Orange
## 5 2021-01-05                   92155                  San Diego San Diego
## 6 2021-01-05                   92259                   Imperial  Imperial
##   vaccine_equity_metric_quartile                vem_source
## 1                              2 Healthy Places Index Score
## 2                              3 Healthy Places Index Score
## 3                              1 Healthy Places Index Score
## 4                              3 Healthy Places Index Score
## 5                             NA            No VEM Assigned
## 6                              1    CDPH-Derived ZCTA Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1               76455.9                84200                       19
## 2               44238.8                47883                       NA
## 3                7098.5                 8026                       NA
## 4               16027.4                16053                       NA
## 5                 456.0                  456                       NA
## 6                 119.0                  121                       NA
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                         1282                                0.000226
## 2                           NA                                      NA
## 3                           NA                                      NA
## 4                           NA                                      NA
## 5                           NA                                      NA
## 6                           NA                                      NA
##   percent_of_population_partially_vaccinated
## 1                                   0.015226
## 2                                         NA
## 3                                         NA
## 4                                         NA
## 5                                         NA
## 6                                         NA
##   percent_of_population_with_1_plus_dose
## 1                                0.015452
## 2                                      NA
```

```
## 3                                       NA
## 4                                       NA
## 5                                       NA
## 6                                       NA
##                                                                      redacted
## 1                                                                          No
## 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?

2021-11-16

We will use **lubridate** package to make life a lot easier when dealing with dates and times

```
library(lubridate)
```

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

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

```
today()
```

```
## [1] "2021-11-23"
```

We make our 'as_of_date' column lubridate format…

```
# Specify that we
vax$as_of_date <- ymd(vax$as_of_date)
```

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

```
## Time difference of 322 days
```

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

```
## Time difference of 7 days
```

Let's quickly look at the data structure using **skim()** function

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

Table 1: Data summary

Name	vax
Number of rows	81144
Number of columns	14
Column type frequency:	
character	4
Date	1
numeric	9
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
local_health_jurisdiction	0	1	0	15	230	62	0
county	0	1	0	15	230	59	0
vem_source	0	1	15	26	0	3	0
redacted	0	1	2	69	0	2	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
as_of_date	0	1	2021-01-05	2021-11-16	2021-06-11	46

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	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
vaccine_equity_metric_quartile	1002	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	
age12_plus_population	0	1.00	18895.04	18993.94	0	1346.95	13685.10	31756.12	88556.7	
age5_plus_population	0	1.00	20875.24	21106.05	0	1460.50	15364.00	34877.00	101902.0	
persons_fully_vaccinated	8256	0.90	9456.49	11498.25	11	506.00	4105.00	15859.00	71078.0	
persons_partially_vaccinated	8256	0.90	1900.61	2113.07	11	200.00	1271.00	2893.00	20185.0	
percent_of_population_fully_vaccinated	8256	0.90	0.42	0.27	0	0.19	0.44	0.62	1.0	
percent_of_population_partially_vaccinated	8256	0.90	0.10	0.10	0	0.06	0.07	0.11	1.0	
percent_of_population_with_1_plus_dose	8256	0.90	0.50	0.26	0	0.30	0.53	0.70	1.0	

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?

8256 missing values

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

10.17%

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

Q9. How many days between the first and last entry in the dataset?

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

```
## Time difference of 315 days
```

315 days between them

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

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

```
## [1] 46
```

46 unique dates

This makes sense because

```
46*7
```

```
## [1] 322
```

We will use **zipcodeR** package to help make sense of the zipcodes

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

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

To calculate distance between two zipcodes:

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

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

pull census data about ZIP code areas (including median household income etc.):

```
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>
```

reverse_zipcode() pulls census data later on for any or all ZIP code areas we might be interested in

Focus on San Diego county

```
table(vax$county)
```

```
##
##                      Alameda          Alpine          Amador           Butte
##            230          2254              46             552             828
##       Calaveras          Colusa    Contra Costa       Del Norte       El Dorado
##            828             322            1978             184            1012
##          Fresno           Glenn       Humboldt        Imperial            Inyo
##           2530             276            1610             690             460
##           Kern           Kings            Lake          Lassen     Los Angeles
##           2254             322             644             598           13340
##         Madera          Marin       Mariposa       Mendocino          Merced
##            552            1288             368            1196             874
##          Modoc           Mono       Monterey           Napa          Nevada
##            506             322            1288             460             552
##         Orange          Placer          Plumas       Riverside      Sacramento
##           4048            1334             736            3220            2484
##      San Benito  San Bernardino       San Diego   San Francisco     San Joaquin
##            184            4094            4922            1242            1472
## San Luis Obispo       San Mateo   Santa Barbara     Santa Clara      Santa Cruz
##           1012            1334            1058            2668             782
##         Shasta          Sierra       Siskiyou          Solano          Sonoma
##           1196             322             966             690            1656
##      Stanislaus          Sutter          Tehama         Trinity          Tulare
##           1104             414             598             598            1518
##        Tuolumne         Ventura           Yolo            Yuba
##            598            1242             782             506
```

We will subset with base R

```
inds <- vax$county == "San Diego"

head(vax[inds,])
```

```
##    as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 5  2021-01-05                   92155                San Diego San Diego
## 14 2021-01-05                   92147                San Diego San Diego
## 16 2021-01-05                   92124                San Diego San Diego
## 24 2021-01-05                   92145                San Diego San Diego
## 34 2021-01-05                   91935                San Diego San Diego
## 36 2021-01-05                   92102                San Diego San Diego
##    vaccine_equity_metric_quartile               vem_source
## 5                              NA           No VEM Assigned
## 14                             NA           No VEM Assigned
## 16                              3 Healthy Places Index Score
## 24                             NA           No VEM Assigned
## 34                              3 Healthy Places Index Score
## 36                              1 Healthy Places Index Score
##    age12_plus_population age5_plus_population persons_fully_vaccinated
## 5                  456.0                  456                       NA
## 14                 518.0                  518                       NA
## 16               25422.4                29040                       29
## 24                1603.5                 1821                       NA
## 34                7390.0                 8101                       NA
## 36               37042.3                41033                       29
```

```
##    persons_partially_vaccinated percent_of_population_fully_vaccinated
## 5                            NA                                      NA
## 14                           NA                                      NA
## 16                          573                                0.000999
## 24                           NA                                      NA
## 34                           NA                                      NA
## 36                         1495                                0.000707
##    percent_of_population_partially_vaccinated
## 5                                          NA
## 14                                         NA
## 16                                   0.019731
## 24                                         NA
## 34                                         NA
## 36                                   0.036434
##    percent_of_population_with_1_plus_dose
## 5                                      NA
## 14                                     NA
## 16                               0.020730
## 24                                     NA
## 34                                     NA
## 36                               0.037141
##                                                                    redacted
## 5  Information redacted in accordance with CA state privacy requirements
## 14 Information redacted in accordance with CA state privacy requirements
## 16                                                                      No
## 24 Information redacted in accordance with CA state privacy requirements
## 34 Information redacted in accordance with CA state privacy requirements
## 36                                                                      No
```

Can be difficult to do this way, try **dplyr** package and its **filter()** function:

```
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] 4922
```

How many entries are there for San Diego county?

```
nrow(sd)
```

```
## [1] 4922
```

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

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

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

107

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

```
ind <- which.max(sd$age12_plus_population)
sd[ind,]
```

```
##    as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 23 2021-01-05                   92154                San Diego San Diego
##    vaccine_equity_metric_quartile                vem_source
## 23                              2 Healthy Places Index Score
##    age12_plus_population age5_plus_population persons_fully_vaccinated
## 23               76365.2                82971                       32
##    persons_partially_vaccinated percent_of_population_fully_vaccinated
## 23                         1336                                0.000386
##    percent_of_population_partially_vaccinated
## 23                                   0.016102
##    percent_of_population_with_1_plus_dose redacted
## 23                                0.016488       No
```

92154

What is the population in the 92037 ZIP Code area?

```
filter(sd, zip_code_tabulation_area == "92037")[1,]
```

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 1 2021-01-05                  92037                San Diego San Diego
##   vaccine_equity_metric_quartile                vem_source
## 1                              4 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1              33675.6                36144                       44
```

```
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                         1265                                0.001217
##   percent_of_population_partially_vaccinated
## 1                                   0.034999
##   percent_of_population_with_1_plus_dose redacted
## 1                                0.036216       No
```

36144 are age 5+

Q13. What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2021-11-09"?

```
sd.now <- filter(sd, as_of_date == "2021-11-09")

mean(sd.now$percent_of_population_fully_vaccinated, na.rm=TRUE)
```

```
## [1] 0.6727567
```

67.3% are fully vaccinated

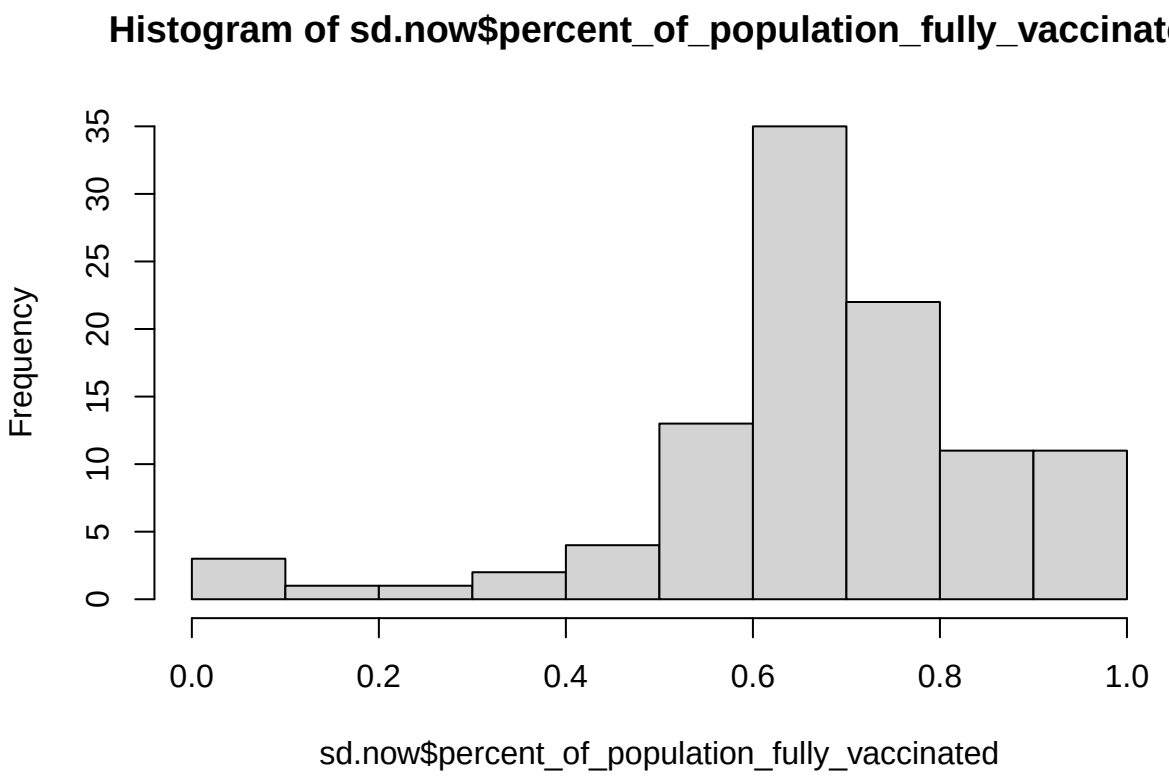
```
summary(sd.now$percent_of_population_fully_vaccinated)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.01017 0.60776 0.67700 0.67276 0.76164 1.00000       4
```

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 "2021-11-09"?

Make a histogram using base R

```
hist(sd.now$percent_of_population_fully_vaccinated)
```

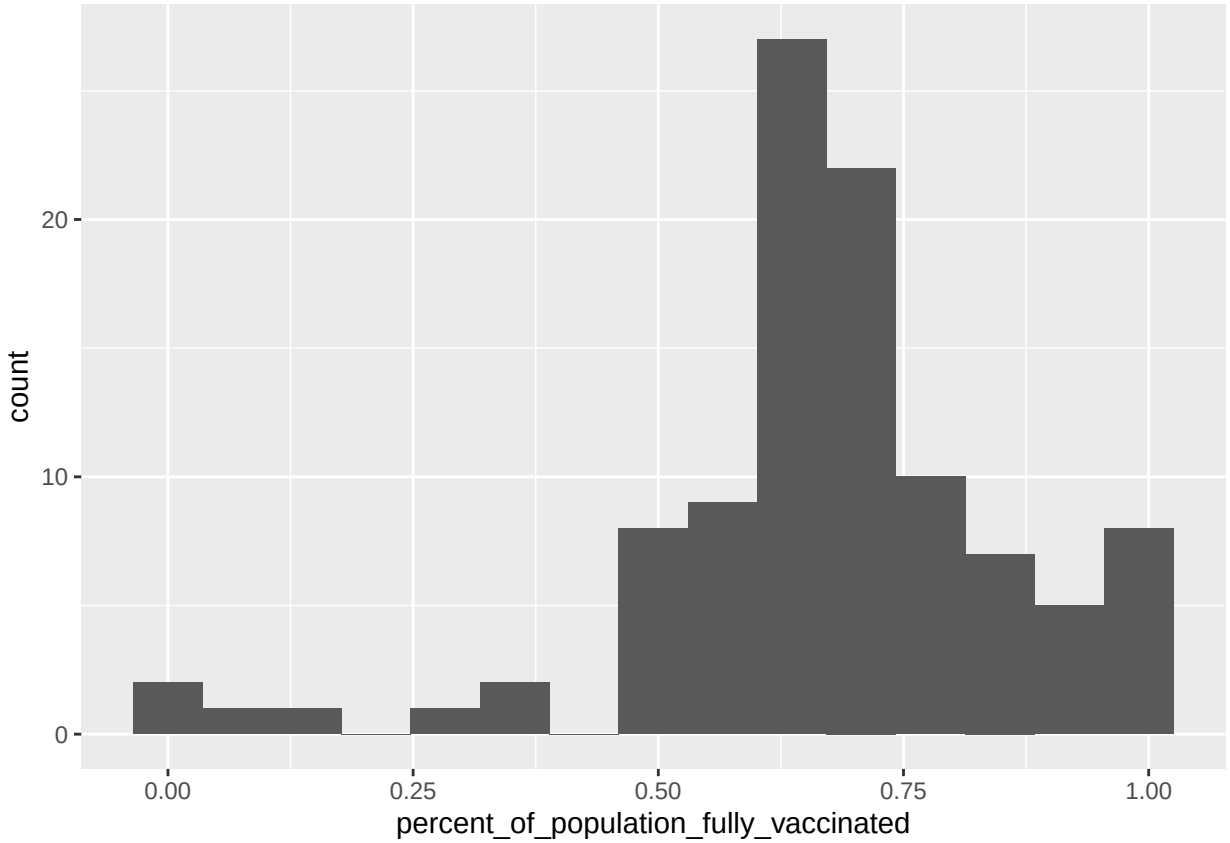


Using ggplot instead:

```
library(ggplot2)

ggplot(sd.now) +aes(percent_of_population_fully_vaccinated) + geom_histogram(bins=15)
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



What about 92037 - UCSD/La Jolla area?

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

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 1 2021-01-05                   92037                San Diego San Diego
## 2 2021-01-12                   92037                San Diego San Diego
## 3 2021-01-19                   92037                San Diego San Diego
## 4 2021-01-26                   92037                San Diego San Diego
## 5 2021-02-02                   92037                San Diego San Diego
## 6 2021-02-09                   92037                San Diego San Diego
##   vaccine_equity_metric_quartile                vem_source
## 1                              4 Healthy Places Index Score
## 2                              4 Healthy Places Index Score
## 3                              4 Healthy Places Index Score
## 4                              4 Healthy Places Index Score
## 5                              4 Healthy Places Index Score
## 6                              4 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1               33675.6                36144                       44
## 2               33675.6                36144                      470
## 3               33675.6                36144                      730
## 4               33675.6                36144                     1079
## 5               33675.6                36144                     1616
## 6               33675.6                36144                     2222
```

```
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                         1265                                0.001217
## 2                         1565                                0.013004
## 3                         3505                                0.020197
## 4                         6197                                0.029853
## 5                         8388                                0.044710
## 6                         9634                                0.061476
##   percent_of_population_partially_vaccinated
## 1                                   0.034999
## 2                                   0.043299
## 3                                   0.096973
## 4                                   0.171453
## 5                                   0.232072
## 6                                   0.266545
##   percent_of_population_with_1_plus_dose redacted
## 1                               0.036216       No
## 2                               0.056303       No
## 3                               0.117170       No
## 4                               0.201306       No
## 5                               0.276782       No
## 6                               0.328021       No
```

```
ucsd[1,]$age5_plus_population
```

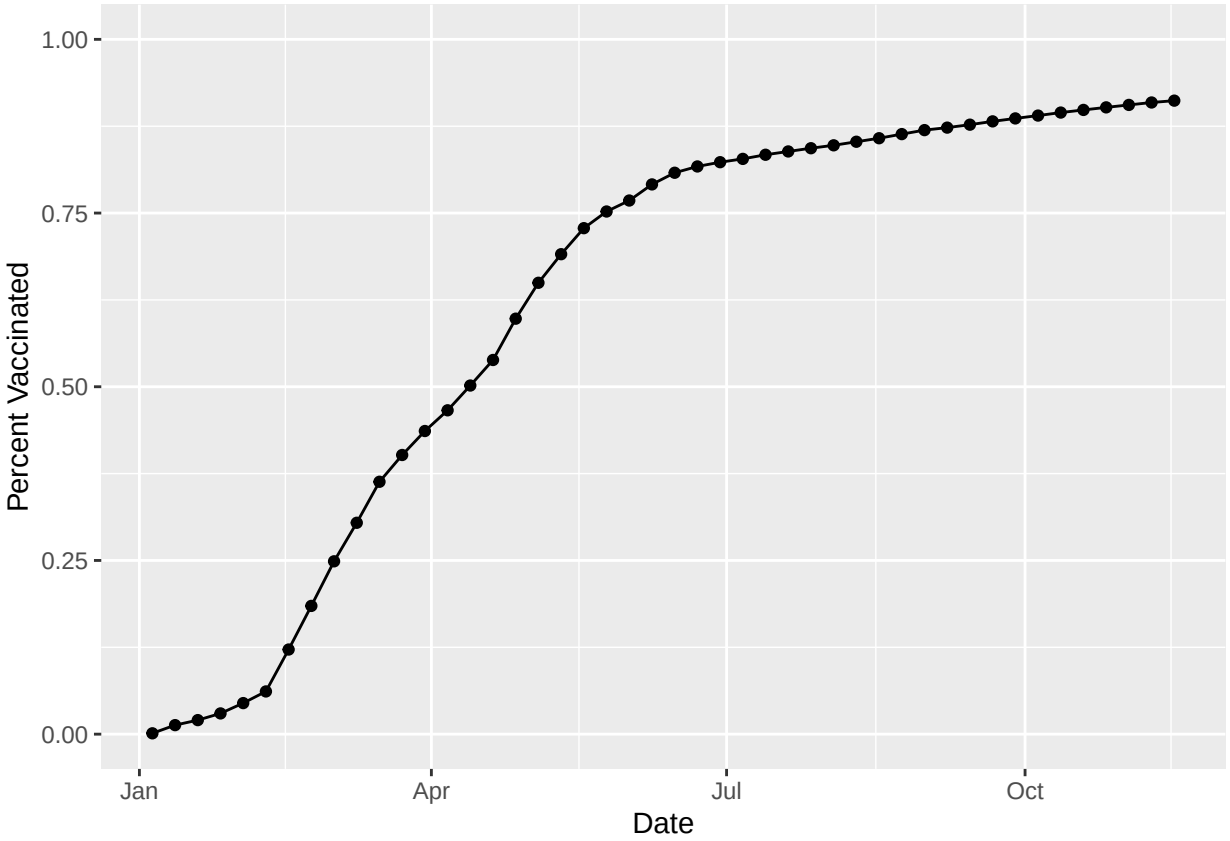
```
## [1] 36144
```

population above age 5: 36,144

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

```
library(ggplot2)

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 92037 to other 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 "2021-11-16".

First subset the full 'vax' dataset to include only zipcode areas with a population as large as 92037:

```
# Subset to all CA areas with a population as large as 92037
vax.36.all <- filter(vax, age5_plus_population > 36144&
                as_of_date == "2021-11-16")

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

```
nrow(vax.36.all)
```

```
## [1] 411
```

How many unique zip codes have a pop. as large as 92037?

```
length(unique(vax.36.all$zip_code_tabulation_area))
```

```
## [1] 411
```

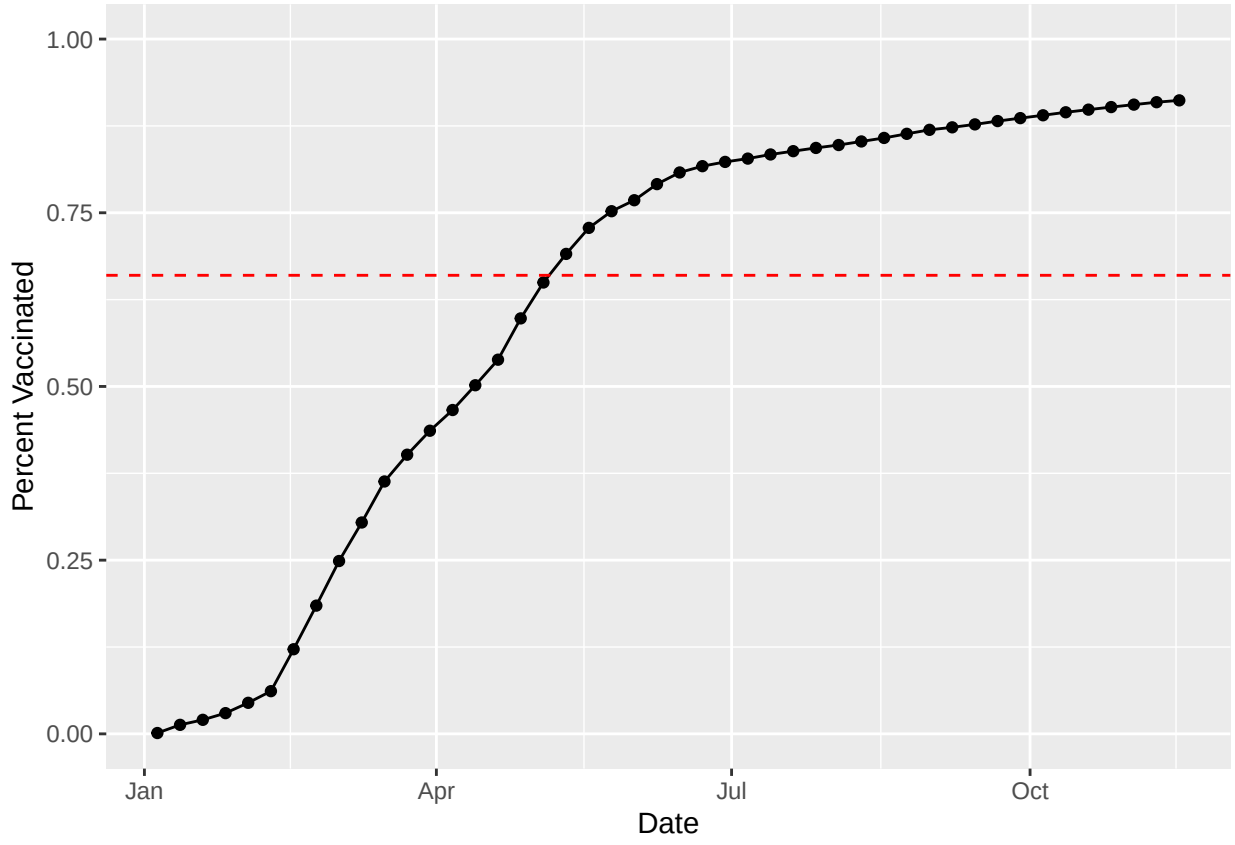
Q16. Calculate the mean "Percent of Population Fully Vaccinated" for ZIP code areas with a population as large as 92037 (La Jolla) as_of_date "2021-11-16". Add this as a straight horizontal line to your plot from above with the geom_hline() function?

```
mean(vax.36.all$percent_of_population_fully_vaccinated, na.rm=TRUE)
```

```
## [1] 0.6629812
```

0.66

```
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=0.66, col="red", linetype="dashed")
```



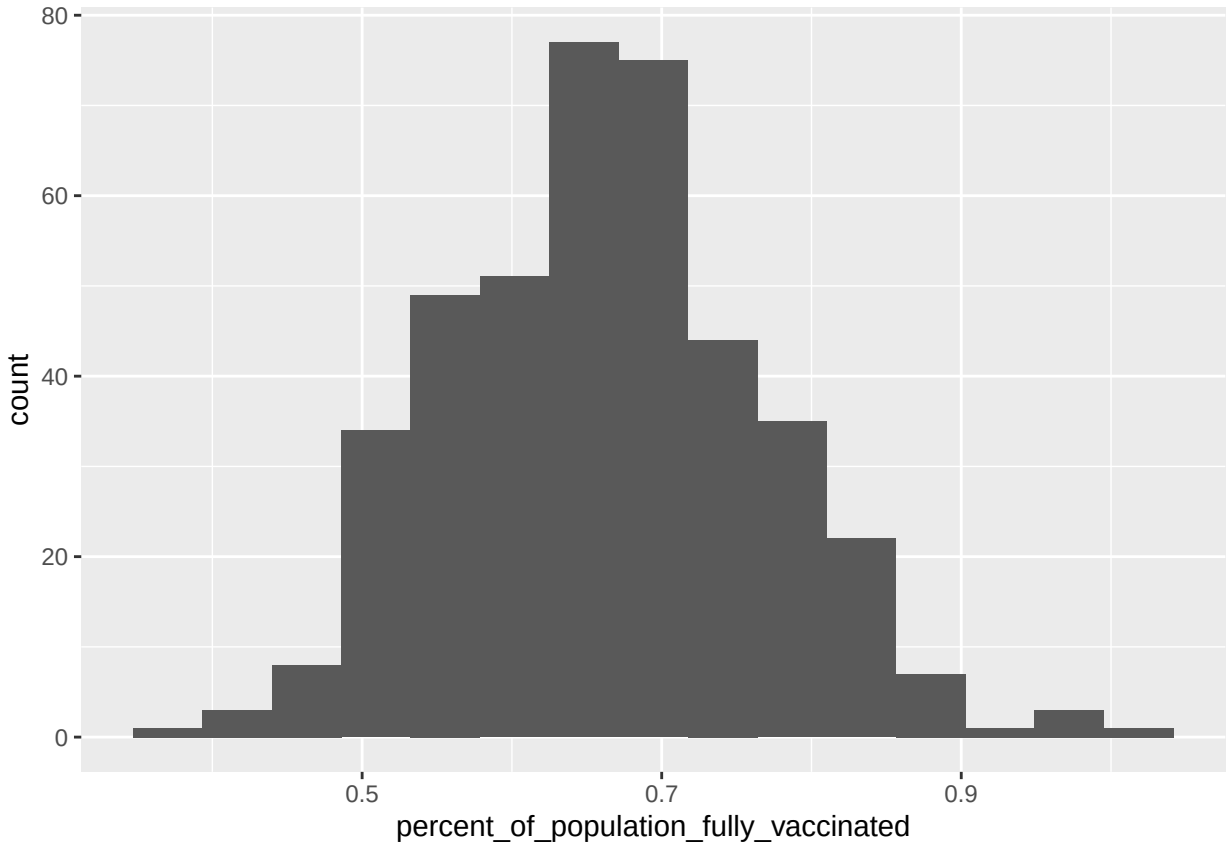
Q17. What is the 6 number summary (Min, 1st Qu., Median, Mean, 3rd Qu., and Max) of the "Percent of Population Fully Vaccinated" values for ZIP code areas with a population as large as 92037 (La Jolla) as_of_date "2021-11-16"?

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3519  0.5891  0.6649  0.6630  0.7286  1.0000
```

Q18. Using ggplot generate a histogram of this data.

```
ggplot(vax.36.all) +aes(percent_of_population_fully_vaccinated) + geom_histogram(bins=15)
```



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 == "2021-11-16") %>%
  filter(zip_code_tabulation_area=="92040") %>%
  select(percent_of_population_fully_vaccinated)
```

```
##   percent_of_population_fully_vaccinated
## 1                               0.520463
```

52%, less than average

```
vax %>% filter(as_of_date == "2021-11-16") %>%
  filter(zip_code_tabulation_area=="92109") %>%
  select(percent_of_population_fully_vaccinated)
```

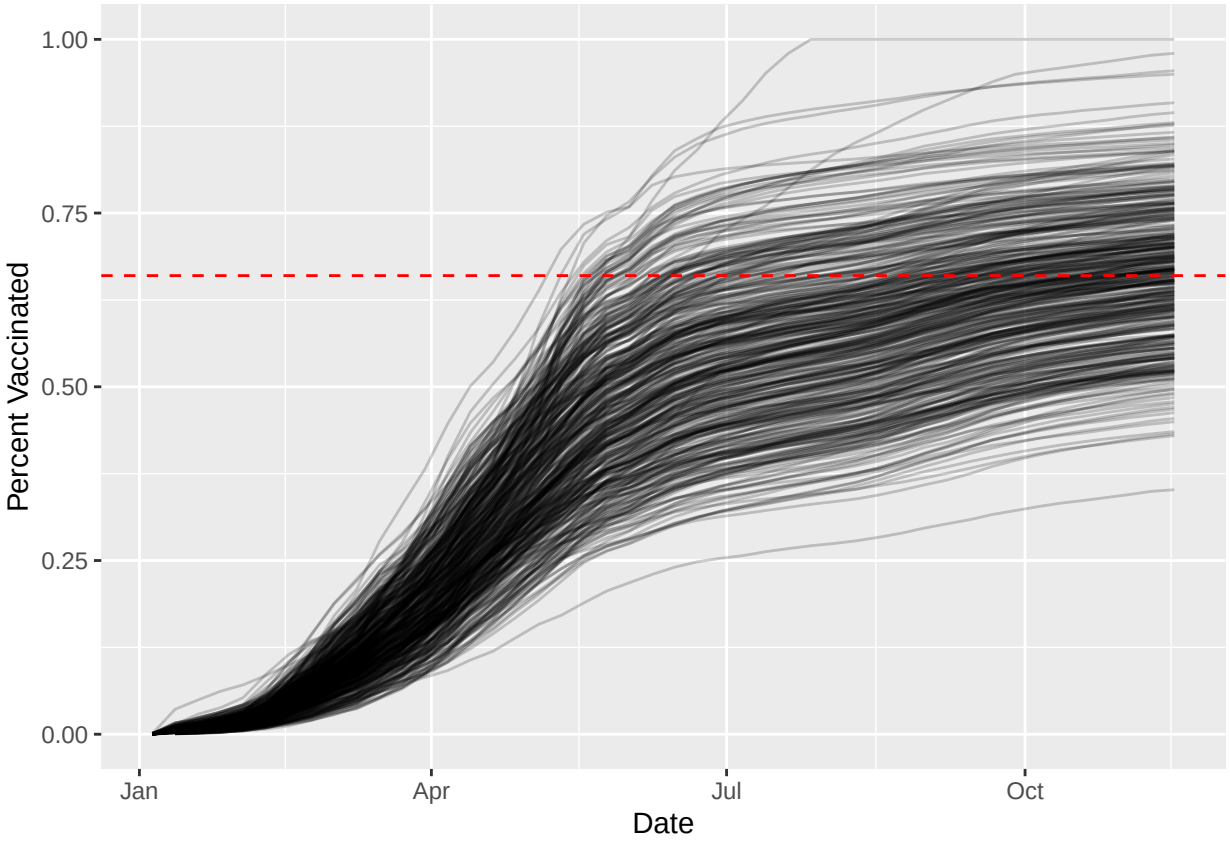
```
##   percent_of_population_fully_vaccinated
## 1                               0.687763
```

68%, above 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.

```
ggplot(vax.36) + aes(as_of_date,
      percent_of_population_fully_vaccinated,
      group=zip_code_tabulation_area) + geom_line(alpha=0.2) + geom_hline(yintercept = 0.66, col="red",
```

```
## Warning: Removed 180 row(s) containing missing values (geom_path).
```



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

It makes me nervous being around people that may have traveled far and been around big groups of people. I see how other people are not as careful as I am and it makes me uneasy.