# class17

Anel A15426506

11/28/2021

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

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction         county
## 1 2021-01-05                    92395            San Bernardino San Bernardino
## 2 2021-01-05                    93206                      Kern           Kern
## 3 2021-01-05                    91006               Los Angeles    Los Angeles
## 4 2021-01-05                    91901                 San Diego      San Diego
## 5 2021-01-05                    92230                 Riverside      Riverside
## 6 2021-01-05                    92662                    Orange         Orange
##   vaccine_equity_metric_quartile                 vem_source
## 1                              1 Healthy Places Index Score
## 2                              1 Healthy Places Index Score
## 3                              3 Healthy Places Index Score
## 4                              3 Healthy Places Index Score
## 5                              1 Healthy Places Index Score
## 6                              4 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1               35915.3                40888                       NA
## 2                1237.5                 1521                       NA
## 3               28742.7                31347                       19
## 4               15549.8                16905                       12
## 5                2320.2                 2526                       NA
## 6                2349.5                 2397                       NA
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                           NA                                     NA
## 2                           NA                                     NA
## 3                          873                               0.000606
## 4                          271                               0.000710
## 5                           NA                                     NA
## 6                           NA                                     NA
##   percent_of_population_partially_vaccinated
## 1                                         NA
## 2                                         NA
## 3                                   0.027850
## 4                                   0.016031
## 5                                         NA
## 6                                         NA
##   percent_of_population_with_1_plus_dose
## 1                                     NA
## 2                                     NA
## 3                               0.028456
## 4                               0.016741
## 5                                     NA
## 6                                     NA
##                                                                redacted
## 1 Information redacted in accordance with CA state privacy requirements
## 2 Information redacted in accordance with CA state privacy requirements
## 3                                                                    No
## 4                                                                    No
## 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_vacinated

> 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] "2021-11-29"
```

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 328 days
```

## Time difference of 322 days

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

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

## 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 | 82908 |
| 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 | 235 | 62 | 0 |
| county | 0 | 1 | 0 | 15 | 235 | 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 | 2021-11-23 | 2021-06-15 | 47 |

**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 | 4089 | 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 | 21106.04 | 0 | 1460.50 | 15364.00 | 34877.00 | 101902.0 | |
| persons_fully_vaccinated | 8355 | 0.90 | 9585.35 | 11609.12 | 11 | 516.00 | 4210.00 | 16095.00 | 71219.0 | |
| persons_partially_vaccinated | 8355 | 0.90 | 1894.87 | 2105.55 | 11 | 198.00 | 1269.00 | 2880.00 | 20159.0 | |
| percent_of_population_fully_vaccinated | 8355 | 0.90 | 0.43 | 0.27 | 0 | 0.20 | 0.44 | 0.63 | 1.0 | |
| percent_of_population_partially_vaccinated | 8355 | 0.90 | 0.10 | 0.10 | 0 | 0.06 | 0.07 | 0.11 | 1.0 | |
| percent_of_population_with_1_plus_dose | 8355 | 0.90 | 0.51 | 0.26 | 0 | 0.31 | 0.53 | 0.71 | 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 have passed since the last update of the dataset?

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

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

322 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] 47
```

47 unique dates

The answer makes sense because

```
47*7
```

```
## [1] 329
```

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>
```

##Focus on San Diego County

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

```
## 
##                          Alameda          Alpine          Amador           Butte 
##             235            2303              47             564             846 
##       Calaveras          Colusa    Contra Costa       Del Norte       El Dorado 
##             846             329            2021             188            1034 
##          Fresno           Glenn        Humboldt        Imperial            Inyo 
##            2585             282            1645             705             470 
##            Kern           Kings            Lake          Lassen     Los Angeles 
##            2303             329             658             611           13630 
##          Madera           Marin        Mariposa       Mendocino          Merced 
##             564            1316             376            1222             893 
##           Modoc            Mono        Monterey            Napa          Nevada 
##             517             329            1316             470             564 
##          Orange          Placer          Plumas       Riverside      Sacramento 
##            4136            1363             752            3290            2538
##      San Benito  San Bernardino       San Diego   San Francisco     San Joaquin
##             188            4183            5029            1269            1504
## San Luis Obispo       San Mateo   Santa Barbara     Santa Clara      Santa Cruz
##            1034            1363            1081            2726             799
##          Shasta          Sierra        Siskiyou          Solano          Sonoma
##            1222             329             987             705            1692
##      Stanislaus          Sutter          Tehama         Trinity          Tulare
##            1128             423             611             611            1551
##        Tuolumne         Ventura            Yolo            Yuba
##             611            1269             799             517
```

```
inds <- vax$county == "San Diego"
head(vax[inds,])
```

```
##    as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 4  2021-01-05                    91901                 San Diego San Diego
## 14 2021-01-05                    91902                 San Diego San Diego
## 21 2021-01-05                    92011                 San Diego San Diego
## 22 2021-01-05                    92055                 San Diego San Diego
## 25 2021-01-05                    92067                 San Diego San Diego
## 33 2021-01-05                    92081                 San Diego San Diego
##    vaccine_equity_metric_quartile                 vem_source
## 4                               3 Healthy Places Index Score
## 14                              4 Healthy Places Index Score
## 21                              4 Healthy Places Index Score
## 22                              3    CDPH-Derived ZCTA Score
## 25                              4 Healthy Places Index Score
## 33                              2 Healthy Places Index Score
##    age12_plus_population age5_plus_population persons_fully_vaccinated
## 4                15549.8                16905                       12
## 14               16620.7                18026                       22
## 21               20503.6                23247                       NA
## 22               11548.0                11654                       NA
## 25                6973.9                 7480                       11
## 33               25558.0                27632                       14
##    persons_partially_vaccinated percent_of_population_fully_vaccinated
## 4                           271                               0.000710
## 14                          374                               0.001220
## 21                           NA                                     NA
## 22                           NA                                     NA
## 25                          241                               0.001471
## 33                          346                               0.000507
##    percent_of_population_partially_vaccinated
## 4                                    0.016031
## 14                                   0.020748
## 21                                         NA
## 22                                         NA
## 25                                   0.032219
## 33                                   0.012522
##    percent_of_population_with_1_plus_dose
## 4                                0.016741
## 14                               0.021968
## 21                                     NA
## 22                                     NA
## 25                               0.033690
## 33                               0.013029
##                                                                  redacted
## 4                                                                      No
## 14                                                                     No
## 21 Information redacted in accordance with CA state privacy requirements
## 22 Information redacted in accordance with CA state privacy requirements
## 25                                                                     No
## 33                                                                     No
```

```
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] 5029
```

How many entries are there for San Diego county?

```
nrow(sd)
```

```
## [1] 5029
```

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

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

```
## [1] 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
## 60 2021-01-05                    92154                 San Diego San Diego
##    vaccine_equity_metric_quartile                 vem_source
## 60                              2 Healthy Places Index Score
##    age12_plus_population age5_plus_population persons_fully_vaccinated
## 60               76365.2                82971                       33
##    persons_partially_vaccinated percent_of_population_fully_vaccinated
## 60                         1341                               0.000398
##    percent_of_population_partially_vaccinated
## 60                                   0.016162
##    percent_of_population_with_1_plus_dose redacted
## 60                                0.01656       No
```

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                       46
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                         1268                               0.001273
##   percent_of_population_partially_vaccinated
## 1                                   0.035082
##   percent_of_population_with_1_plus_dose redacted
## 1                               0.036355       No
```

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

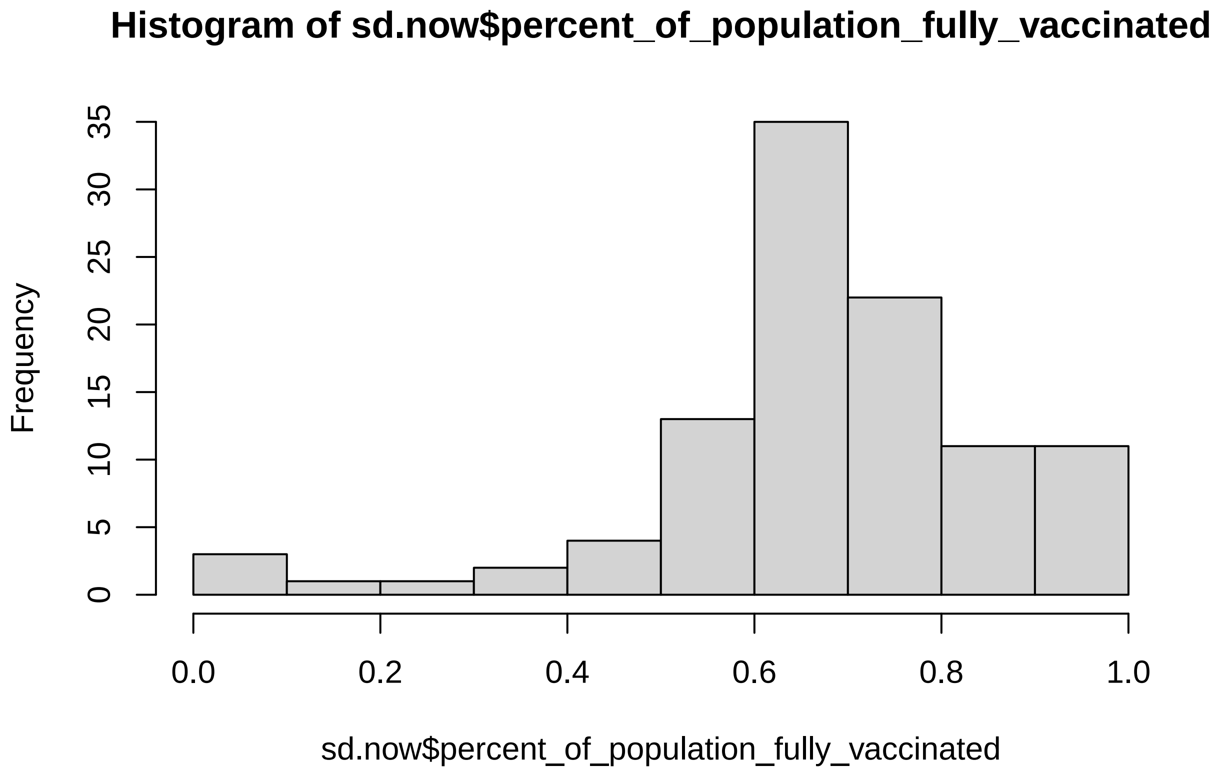
```
sd.now <- filter(sd, as_of_date == "2021-11-09")
mean(sd.now$percent_of_population_fully_vaccinated, na.rm=TRUE)
```

```
## [1] 0.6734714
```

67.3% are fully vaccinated

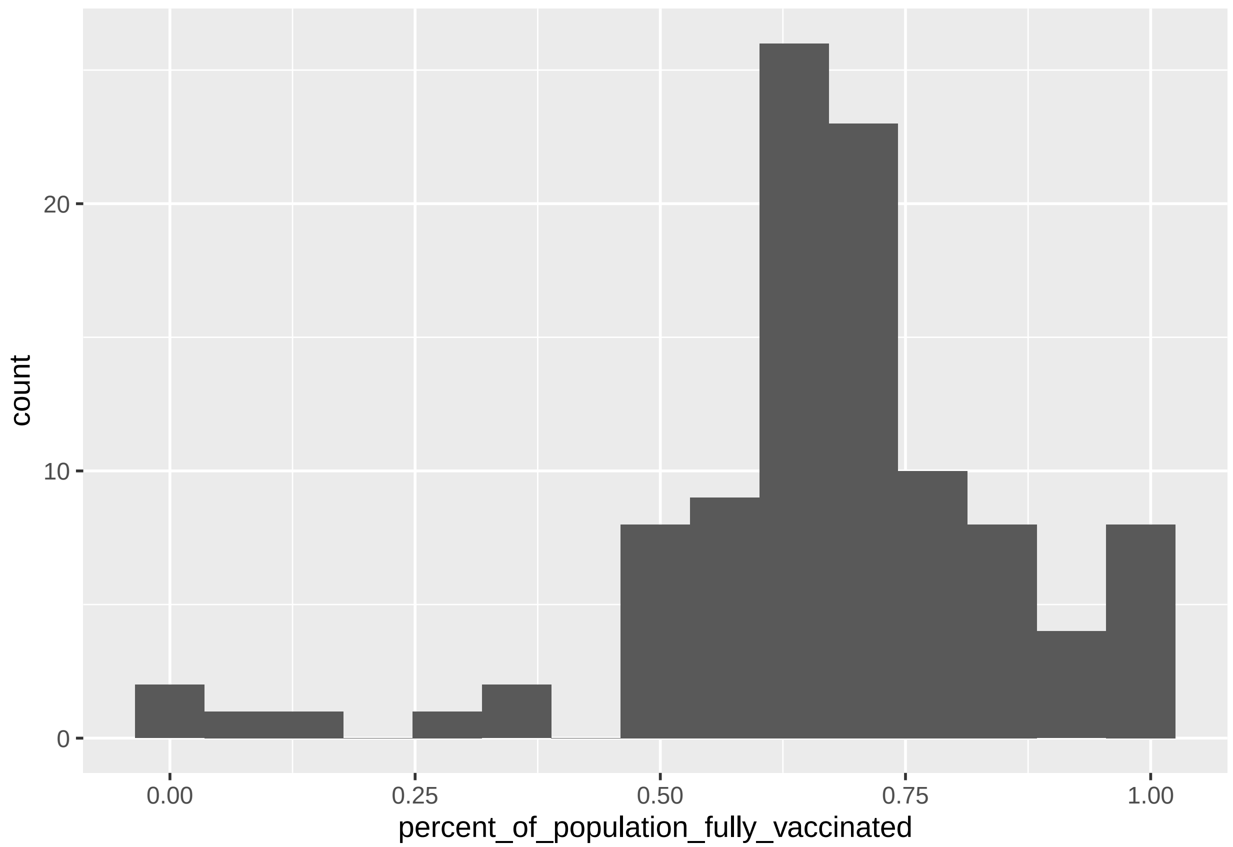
Q14. Using either ggplot or base R graphics make a summary figure that shows the distribution of Percent of Population Fully Vaccinated values as of "2021-11-09"?

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



```
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).
```



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

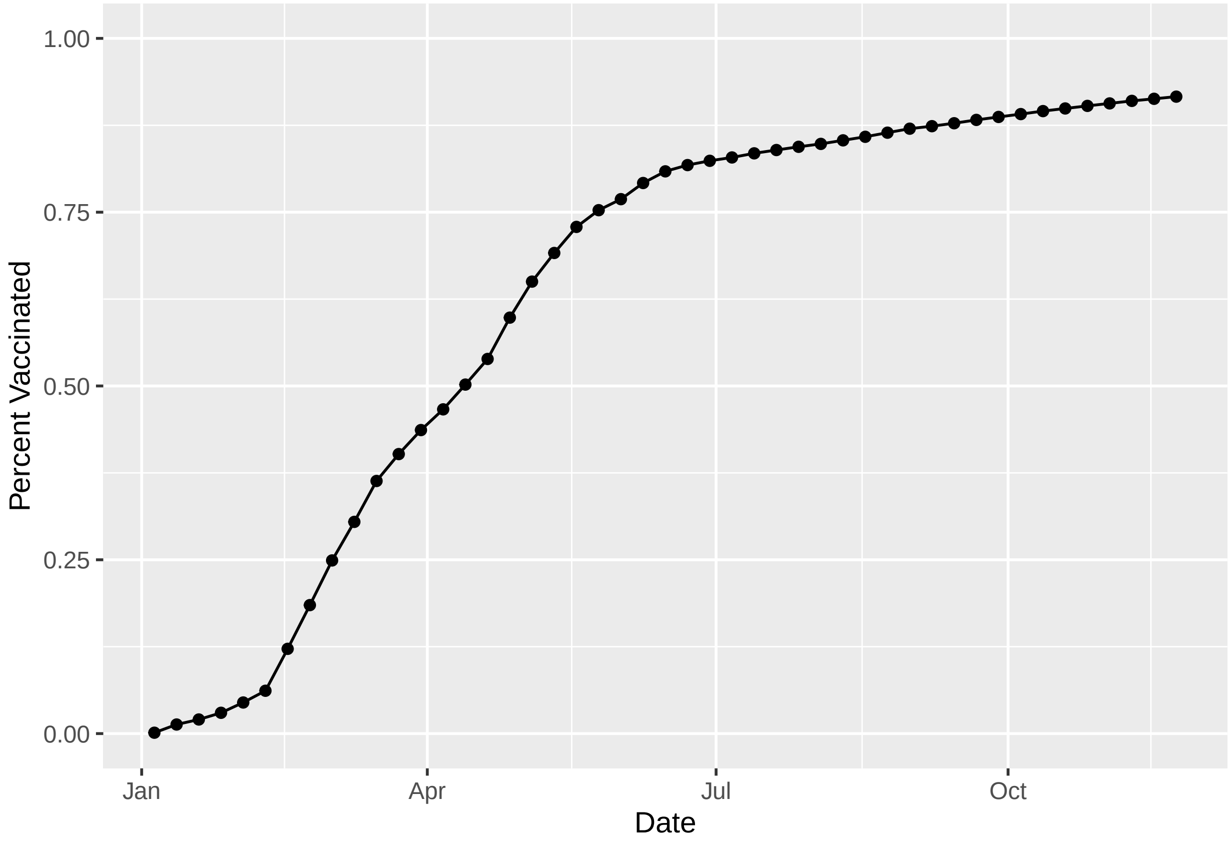
```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 1 2021-01-05                    92037                 San Diego San Diego
## 2 2021-01-12                    92037                 San Diego San Diego
## 3 2021-01-19                    92037                 San Diego 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                       46
## 2               33675.6                36144                      473
## 3               33675.6                36144                      733
## 4               33675.6                36144                     1081
## 5               33675.6                36144                     1617
## 6               33675.6                36144                     2227
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                         1268                               0.001273
## 2                         1569                               0.013087
## 3                         3512                               0.020280
## 4                         6212                               0.029908
## 5                         8408                               0.044738
## 6                         9655                               0.061615
##   percent_of_population_partially_vaccinated
## 1                                   0.035082
## 2                                   0.043410
## 3                                   0.097167
## 4                                   0.171868
## 5                                   0.232625
## 6                                   0.267126
##   percent_of_population_with_1_plus_dose redacted
## 1                               0.036355       No
## 2                               0.056497       No
## 3                               0.117447       No
## 4                               0.201776       No
## 5                               0.277363       No
## 6                               0.328741       No
```

```
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)
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?

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

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

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

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

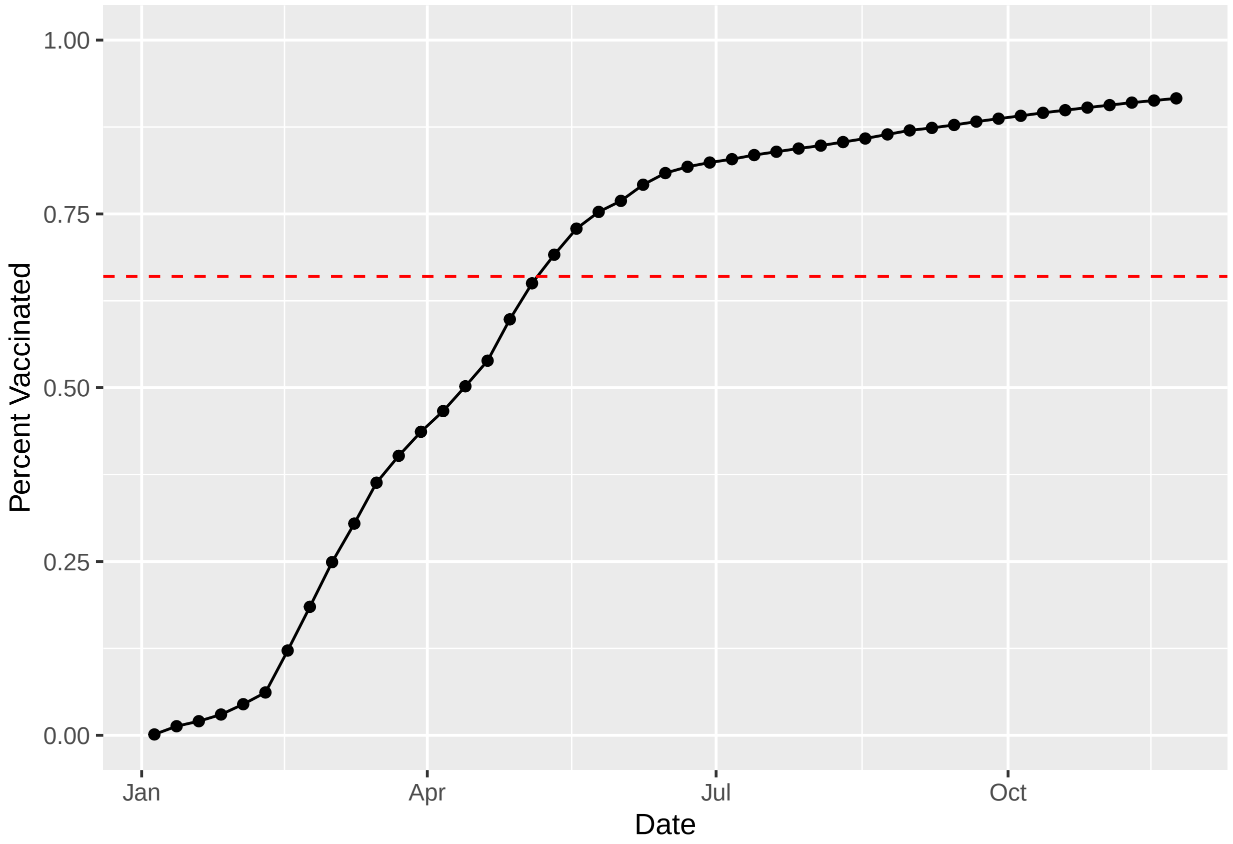
```
## [1] 411
```

> 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] 0.6640413
```

```
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")
```



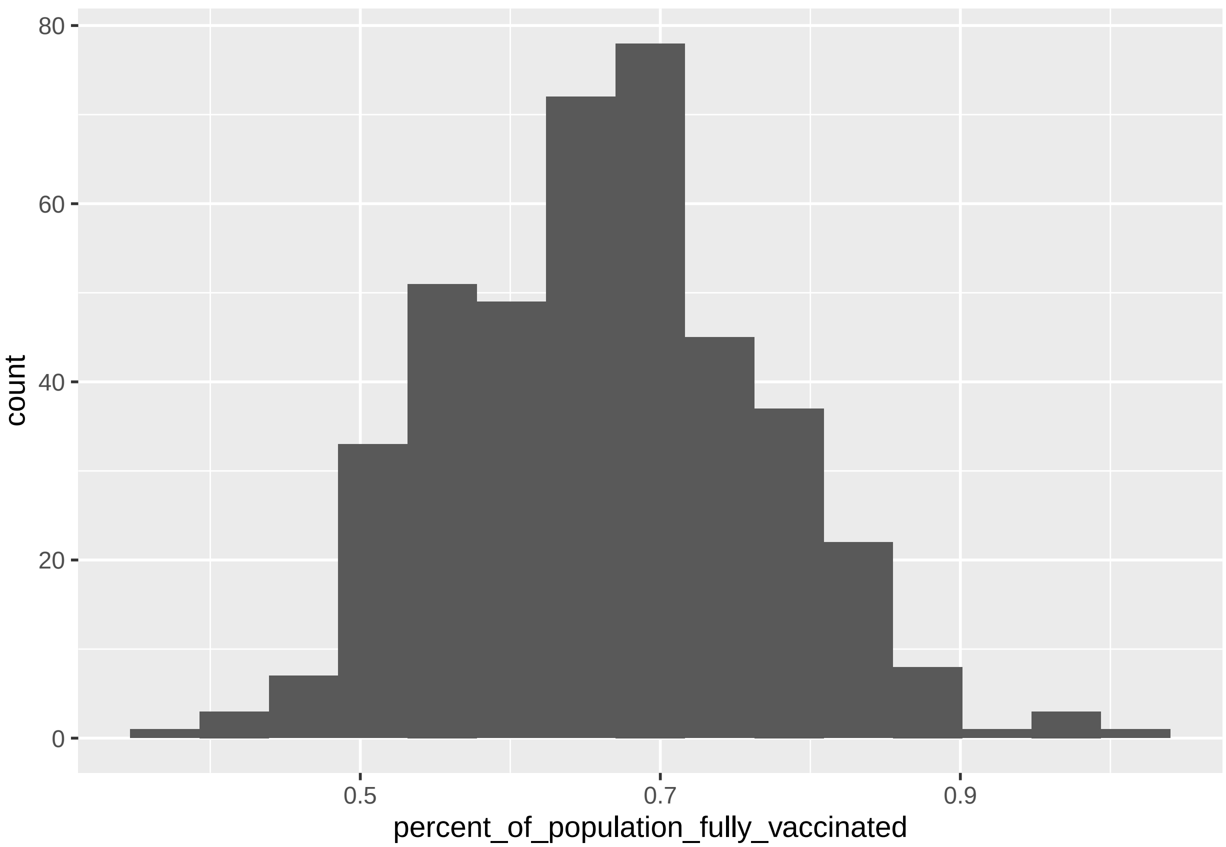
Q17. What is the 6 number summary (Min, 1st Qu., Median, Mean, 3rd Qu., and Max) of the "Percent of Population Fully Vaccinated" values for ZIP code areas with a population as large as 92037 (La Jolla) as_of_date "2021-11-16"?

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3529  0.5905  0.6662  0.6640  0.7298  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                               0.521047
```

52% less than average

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

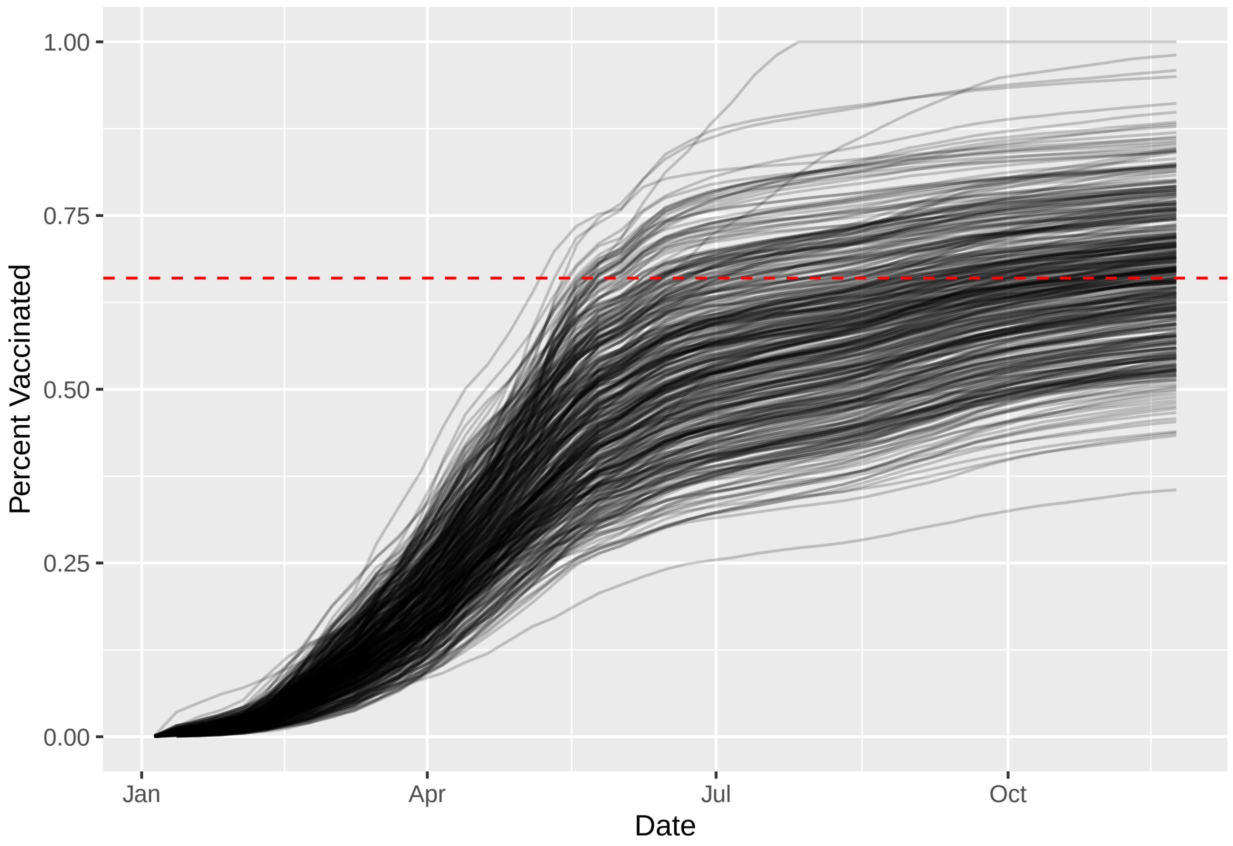
```
##   percent_of_population_fully_vaccinated
## 1                                0.68863
```

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 176 row(s) containing missing values (geom_path).
```



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

I feel a bit uneasy about having class in person because I know most people have traveled and spent their time with big groups.

zz