Class 17: Vaccination rate mini project

Anita Wang (PID: A15567878)

11/23/2021

#Getting Started

```
# Import vaccination data
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?

- The 9th column: "persons_fully_vaccinated"

Q2. What column details the Zip code tabulation area?

- The 2nd column

Q3. What is the earliest date in this dataset?

```
head(vax$as_of_date)
```

```
## [1] "2021-01-05" "2021-01-05" "2021-01-05" "2021-01-05" "2021-01-05"
## [6] "2021-01-05"
```

- 2021-01-05 (January 5th, 2021)

Q4. What is the latest date in this dataset?

```
tail(vax$as_of_date)
```

```
## [1] "2021-11-16" "2021-11-16" "2021-11-16" "2021-11-16" "2021-11-16"
## [6] "2021-11-16"
```

- 2021-11-16 (November 16th, 2021)

As we have done previously, let's call the skim() function from the skimr package to get a quick overview of this dataset:

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

Table 1: Data summary

Name	vax
Number of rows	81144
Number of columns	14
Column type frequency:	
character	5
numeric	9
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	46	0
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: 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	4002	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 columns

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] 8256
```

- 8256 NA values

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

```
per_vax <- sum(is.na(vax$persons_fully_vaccinated)) / sum(vax$persons_fully_vaccinated, na.rm=TRUE)

signif((per_vax *100), 2)
```

```
## [1] 0.0012
```

- 0.0012%

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

- The areas don’t report their vaccination. For example, the military has their own health systems and does not participate in this record keeping.

#Working with dates

##Ensure that the date column is useful

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

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

library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.1.2
```

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

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

```
today()
```

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

Here we make our ~as-of-date‘ column lubridate format

```
# Speciffy that we are using the Year-mont-day format
vax$as_of_date <- ymd(vax$as_of_date)
```

Now we can do math with dates. For example: How many days have passed since the first vaccination reported in this dataset?

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

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

Using the last and the first date value we can now determine how many days the dataset span?

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

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

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

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

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

- 7 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] 46
```

-There are 46 unique dates in the dataset

#Working with ZIP codes

In R we can use the zipcodeR package to make working with these codes easier.

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

```
## Warning: package 'zipcodeR' was built under R version 4.1.2
```

```
geocode_zip('92037')
```

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

Calculate the distance between the centroids of any two ZIP codes in miles, e.g.

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

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

More usefully, we can pull census data about ZIP code areas (including median household income etc.). For example:

```
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>
```

We can use this reverse_zipcode() to pull census data later on for any or all ZIP code areas we might be interested in.

```
# Pull data for all ZIP codes in the dataset
#zipdata <- reverse_zipcode( vax$zip_code_tabulation_area )
```

#Focus on the San Diego area

Let's now focus in on the San Diego County area by restricting ourselves first to vax$county == "San Diego" entries. We have two main choices on how to do this. The first using base R the second using the dplyr package:

Using base R, he code would look like this:

```
# Subset to San Diego county only areas
sd <- vax[ vax$county == "San Diego" , ]
```

Using dplyr the code would look like this:

```
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
```

Using dplyr is often more convenient when we are subsetting across multiple criteria (more advanced subsetting) - for example all San Diego county areas with a population of over 10,000.

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

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

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

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

- There are 107 distinct zip codes

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

```
which.max(sd$age12_plus_population)
```

```
## [1] 23
```

```
sd[23,]
```

```
##    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
```

- zip code: 92154

Using dplyr select all San Diego "county" entries on "as_of_date" "2021-11-16" and use this for the following questions.

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

```
library(dplyr)

sd.now <- filter(sd, as_of_date ==  "2021-11-09")
(mean(sd.now$percent_of_population_fully_vaccinated, na.rm = TRUE))*100
```

```
## [1] 67.27567
```

- 67.28%

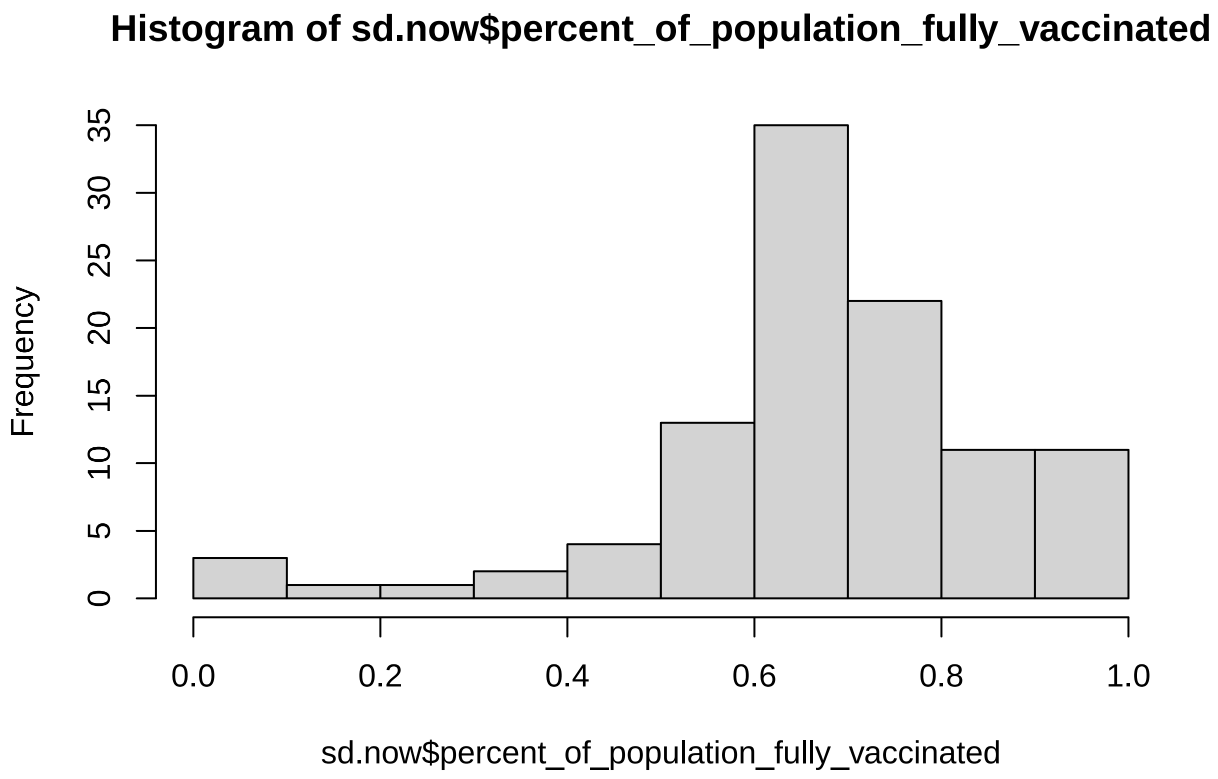
We can also look at the 6-number summary

```
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"?

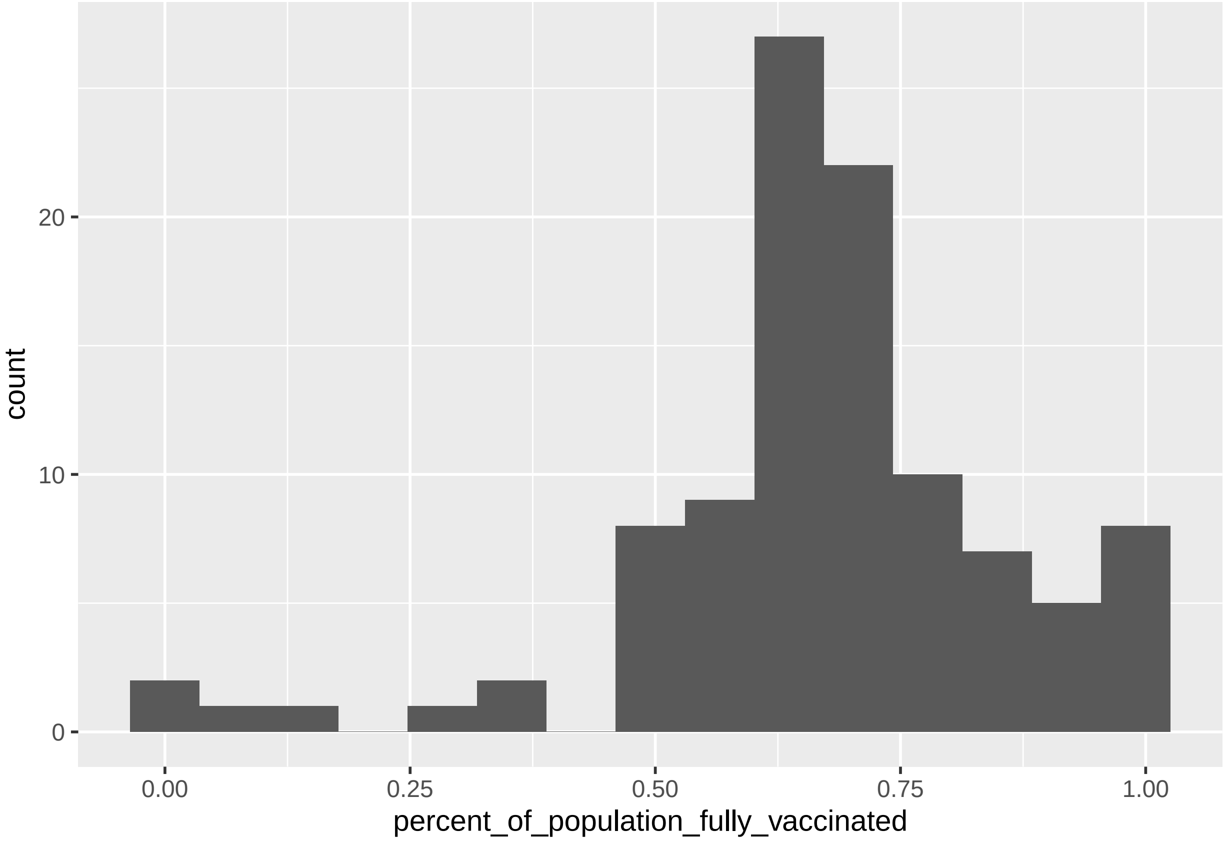
```
hist(sd.now$percent_of_population_fully_vaccinated)
```



or

```
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).
```



#Focus on UCSD/La Jolla

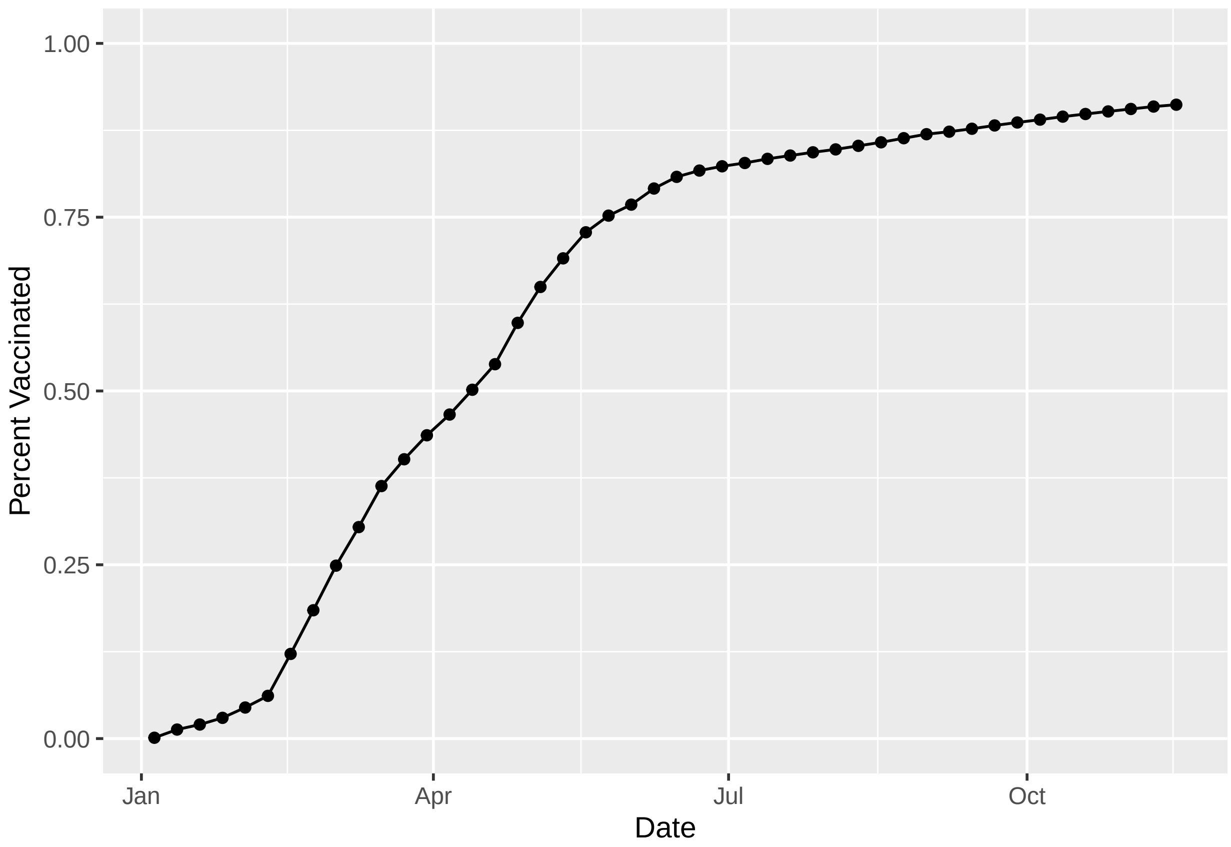
UC San Diego resides in the 92037 ZIP code area and is listed with an age 5+ population size of 36,144.

```
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:

```
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")
```



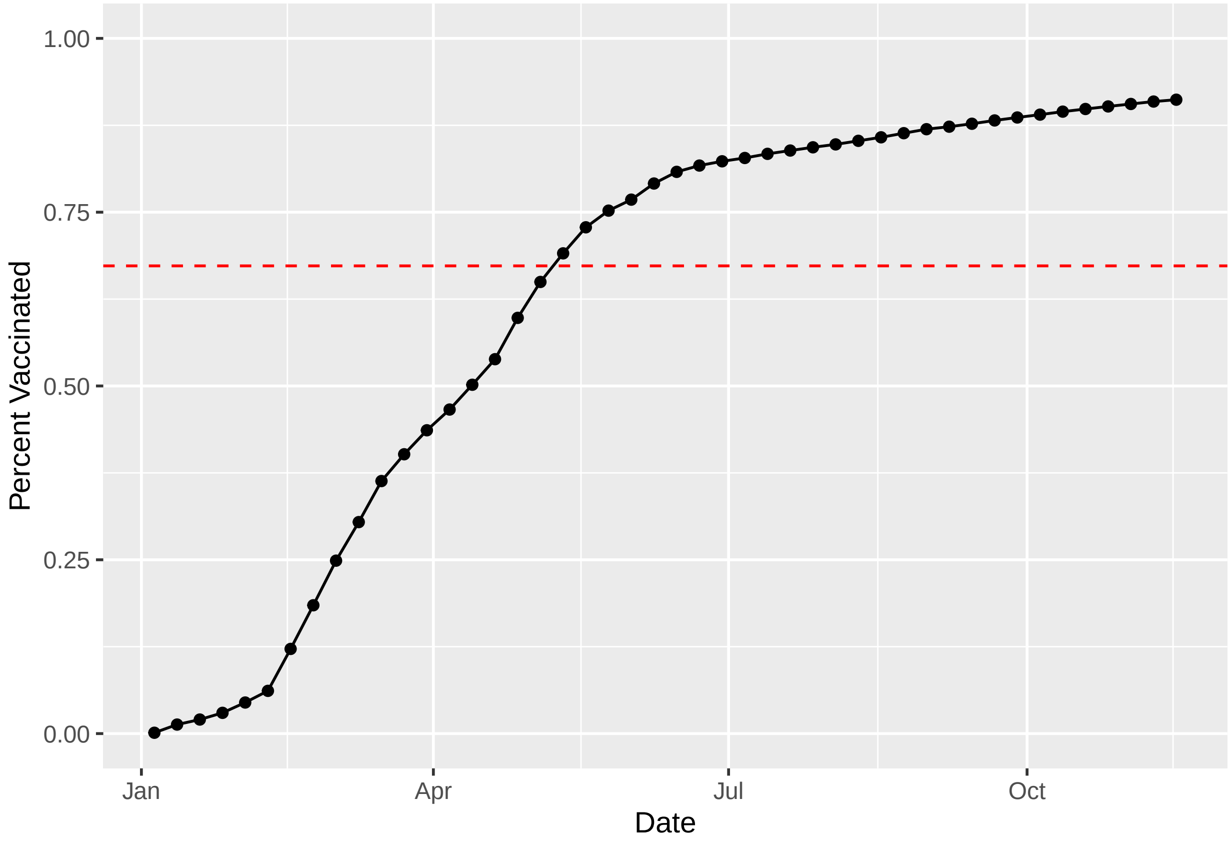
#Comparing 92037 to other similar sized areas?

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

head(vax.36)
```

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction         county
## 1 2021-11-16                    92833                    Orange         Orange
## 2 2021-11-16                    92234                 Riverside      Riverside
## 3 2021-11-16                    92507                 Riverside      Riverside
## 4 2021-11-16                    92555                 Riverside      Riverside
## 5 2021-11-16                    92345            San Bernardino San Bernardino
## 6 2021-11-16                    91306               Los Angeles    Los Angeles
##   vaccine_equity_metric_quartile                 vem_source
## 1                              3 Healthy Places Index Score
## 2                              1 Healthy Places Index Score
## 3                              1 Healthy Places Index Score
## 4                              2 Healthy Places Index Score
## 5                              1 Healthy Places Index Score
## 6                              2 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1               43985.4                48623                    34668
## 2               46401.1                51202                    34191
## 3               51432.5                55253                    31704
## 4               36725.7                41446                    23776
## 5               66047.5                75539                    35332
## 6               42671.1                46573                    31858
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                         3377                               0.712996
## 2                         3966                               0.667767
## 3                         3434                               0.573797
## 4                         2424                               0.573662
## 5                         4428                               0.467732
## 6                         3372                               0.684044
##   percent_of_population_partially_vaccinated
## 1                                   0.069453
## 2                                   0.077458
## 3                                   0.062150
## 4                                   0.058486
## 5                                   0.058619
## 6                                   0.072402
##   percent_of_population_with_1_plus_dose redacted
## 1                               0.782449       No
## 2                               0.745225       No
## 3                               0.635947       No
## 4                               0.632148       No
## 5                               0.526351       No
## 6                               0.756446       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 "2021-11-16". Add this as a straight horizontal line to your plot from above with the geom_hline() function?

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

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") +
  geom_hline(yintercept = 0.6727567, linetype="dashed", col = "red")
```



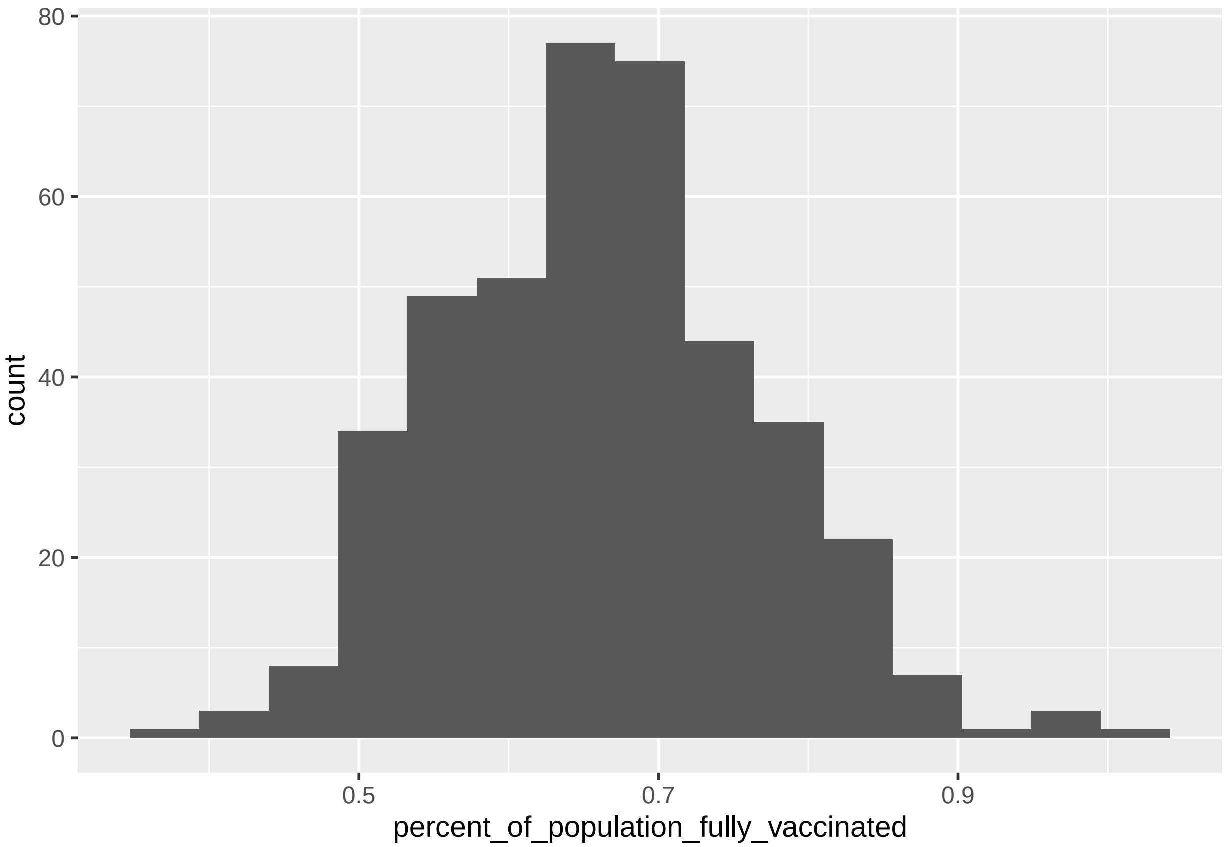
Q17. What is the 6 number summary (Min, 1st Qu., Median, Mean, 3rd Qu., and Max) of the "Percent of Population Fully Vaccinated" values for ZIP code areas with a population as large as 92037 (La Jolla) as_of_date "2021-11-16"?

```
summary(vax.36$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.

```
library(ggplot2)

ggplot(vax.36) +
  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=="92109") %>%
  select(percent_of_population_fully_vaccinated)
```

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

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

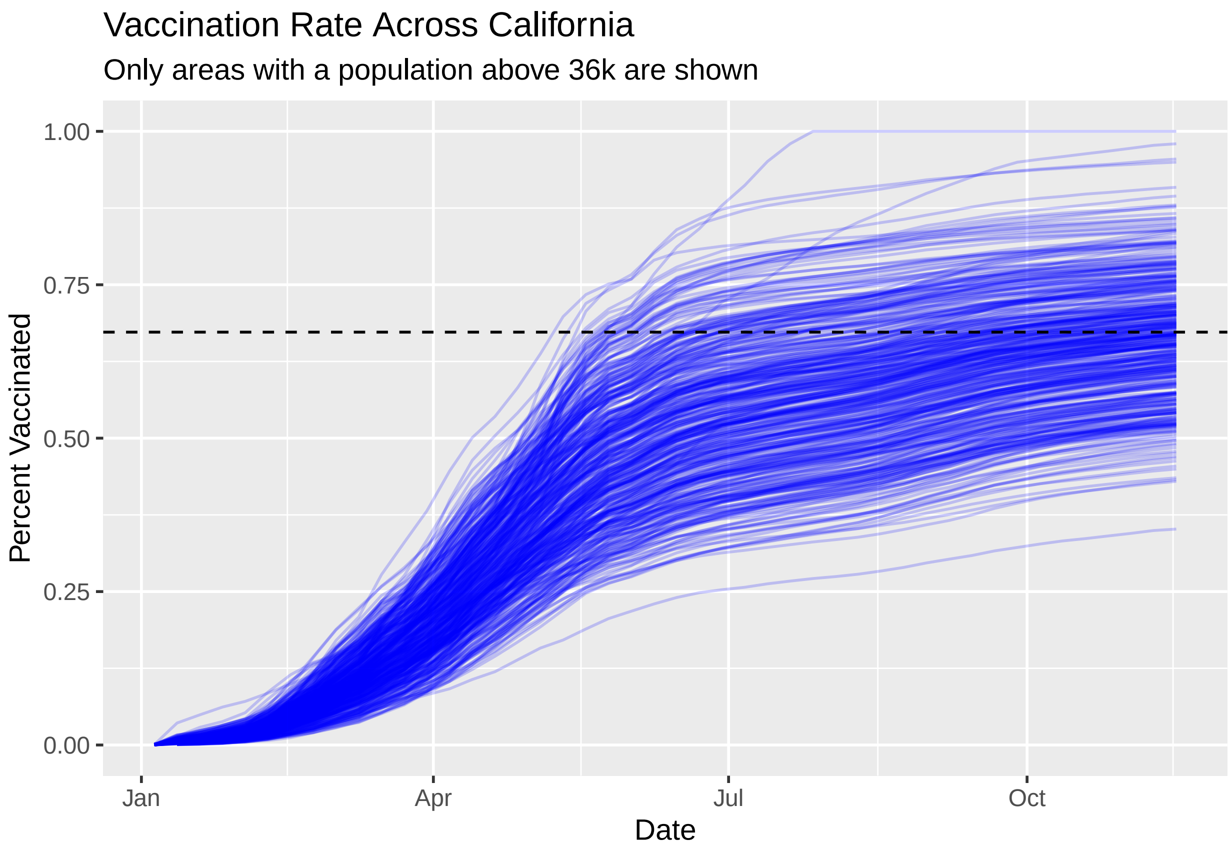
```
##   percent_of_population_fully_vaccinated
## 1                               0.520463
```

- Zip Code 92109 is above average at 68.78% vax where as zip code 92040 is below average at 52.05% vax.

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(c(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 = 0.6727567, linetype="dashed")
```

```
## 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?