class17

Hyeonseok Jang (PID# A59011126)

11/24/2021

##Background The goal of this hands-on mini-project is to examine and compare the Covid-19 vaccination rates around San Diego.

##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                    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
```

> Q. How many entries do we have?

```
nrow(vax)
```

```
## [1] 82908
```

> Q1. What column details the total number of people fully vaccinated?

```
head(vax$persons_fully_vaccinated)
```

```
## [1] NA NA 19 12 NA NA
```

> Q2. What column details the Zip code tabulation area?

```
head(vax$zip_code_tabulation_area)
```

```
## [1] 92395 93206 91006 91901 92230 92662
```

> Q3. What is the earliest date in this dataset?

```
head(sort(vax$as_of_date), 1)
```

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

> Q4. What is the latest date in this dataset?

```
tail(sort(vax$as_of_date), 1)
```

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

We can use the **skimr** package and the ‘skim()’ function to get a quick overview of structure of this dataset.

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

Table 1: Data summary

Name	vax
Number of rows	82908
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	47	0
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: 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 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] 8355
```

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

```r
round(sum(is.na(vax$persons_fully_vaccinated))/nrow(vax)*100, 2)
```

```
## [1] 10.08
```

##Working with dates

Notice that one of the columns is a data column. Working with the data we need **lubridate** package.

```r
##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
```

This will not work because our data column was read as character..

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

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

```r
today()-d[1]
```

```
## Time difference of 333 days
```

I will make the 'as_of_date' column Date format..

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

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

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

```
## Time difference of 11 days
```

> Q. How many days does the dataset span?

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

```
## Time difference of 322 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] 47
```

> Q. How many different ZIP code areas are in this dataset?

```
length(unique(vax$zip_code_tabulation_area))
```

```
## [1] 1764
```

Working with ZIP codes

To work with ZIP codes we can use the **zipcodeR**

```
##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.
```

```
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>
```

##Focus on the San Diego area

We want to subset the full CA 'vax' data down to just San Diego County.

We could do this with base R.

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

Subsetting can get tedious and complicated quickly when you have multiple things we want to subset by.

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

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

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

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

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

More complicated subsetting…

```r
sd.20 <- filter(vax, county=="San Diego",
      age5_plus_population > 20000)
nrow(sd.20)
```

```
## [1] 3055
```

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

```r
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?

```r
sd$zip_code_tabulation_area[which.max(sd$age12_plus_population)]
```

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

> Q13. What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2021-11-23"?

```
sd.now <- filter(sd, as_of_date=="2021-11-23")
mean(sd.now$percent_of_population_fully_vaccinated, na.rm=TRUE)
```

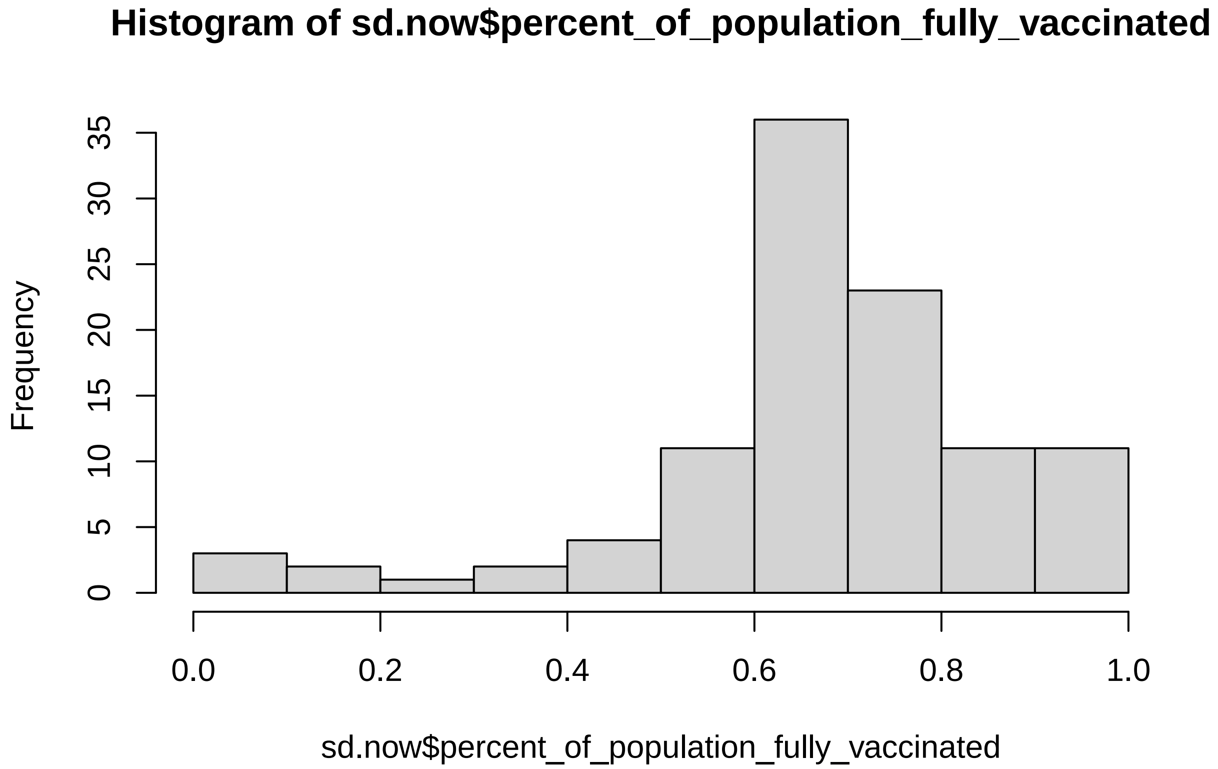
```
## [1] 0.6740001
```

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.01017 0.61301 0.67965 0.67400 0.76932 1.00000       3
```

> 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 "2021-11-23"?

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



This plot above is going to be susceptible to being skewed by ZIP code areas with small population. This will have big effects for just a small number of unvax-ed folks..

##Focus on UCSD/La Jolla

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

> Q. What is the population of the 92037 ZIP code area?

```
unique(lj$age5_plus_population)
```

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

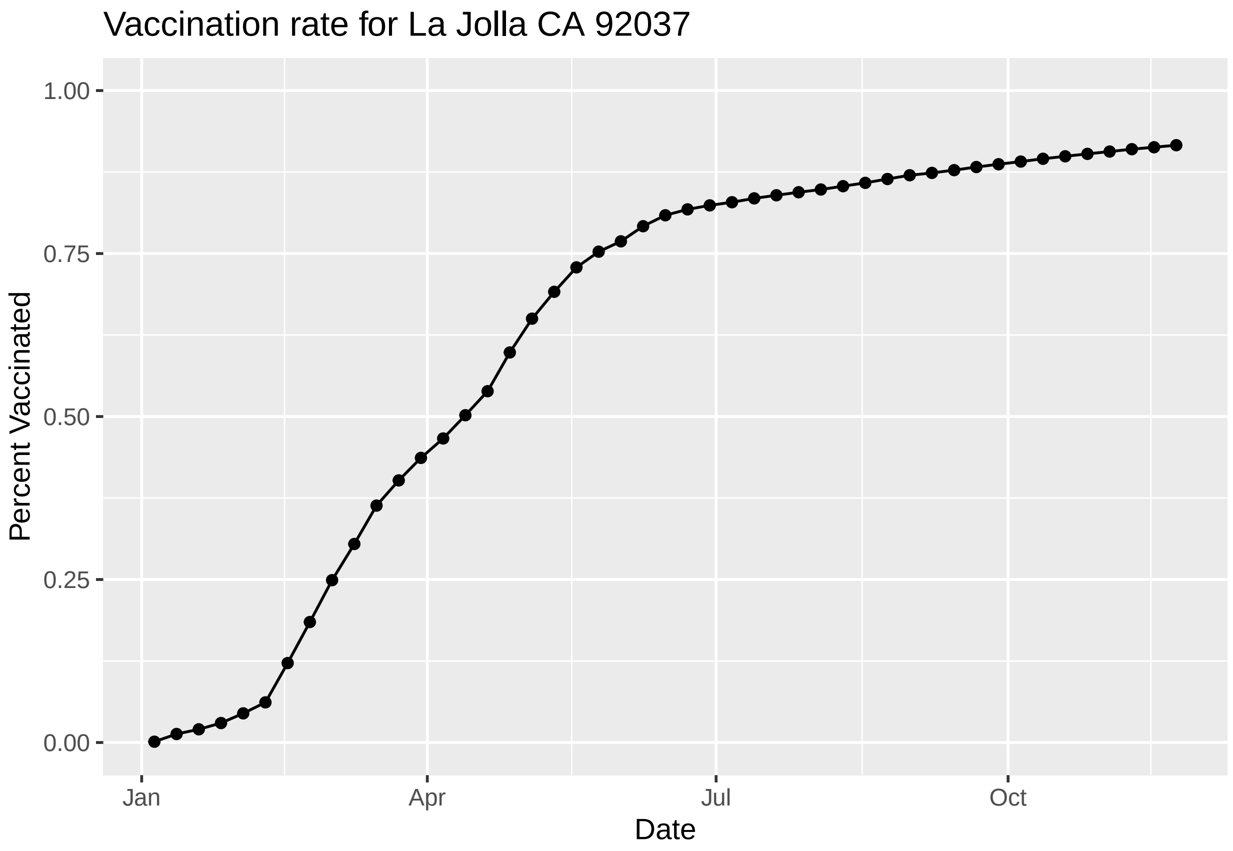
Q. What is the current vaccination value for this UCSD/La Jolla Zip code area?

```
lj2 <- filter(sd.now, zip_code_tabulation_area=="92037")
lj2$percent_of_population_fully_vaccinated
```

```
## [1] 0.916196
```

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

```
library(ggplot2)
ljvaxrate <- ggplot(lj) +
  aes(x=as_of_date,
      y=percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) +
  ylim(c(0, 1)) +
  labs(title="Vaccination rate for La Jolla CA 92037", x="Date", y="Percent Vaccinated")
ljvaxrate
```



##Comparing to similar sized areas Let's make this plot for all San Diego County ZIP cod areas that have a population as least as large as 92037.

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

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 1 2021-01-05                    92058                 San Diego San Diego
## 2 2021-01-05                    92078                 San Diego San Diego
## 3 2021-01-05                    92019                 San Diego San Diego
## 4 2021-01-05                    92117                 San Diego San Diego
## 5 2021-01-05                    92057                 San Diego San Diego
## 6 2021-01-05                    91913                 San Diego San Diego
##   vaccine_equity_metric_quartile                 vem_source
## 1                              1 Healthy Places Index Score
## 2                              3 Healthy Places Index Score
## 3                              3 Healthy Places Index Score
## 4                              3 Healthy Places Index Score
## 5                              2 Healthy Places Index Score
## 6                              3 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1               34956.0                39695                       NA
## 2               41789.5                47476                       37
## 3               37439.4                40464                       25
## 4               50041.6                53839                       42
## 5               51927.0                56906                       22
## 6               43514.7                50461                       37
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                           NA                                     NA
## 2                          688                               0.000779
## 3                          610                               0.000618
## 4                         1143                               0.000780
## 5                          691                               0.000387
## 6                         1993                               0.000733
##   percent_of_population_partially_vaccinated
## 1                                         NA
## 2                                   0.014492
## 3                                   0.015075
## 4                                   0.021230
## 5                                   0.012143
## 6                                   0.039496
##   percent_of_population_with_1_plus_dose
## 1                                     NA
## 2                               0.015271
## 3                               0.015693
## 4                               0.022010
## 5                               0.012530
## 6                               0.040229
##                                                                redacted
## 1 Information redacted in accordance with CA state privacy requirements
## 2                                                                    No
## 3                                                                    No
## 4                                                                    No
## 5                                                                    No
## 6                                                                    No
```

Q. How many ZIP code areas in San Diego county have a population larger than 92037?

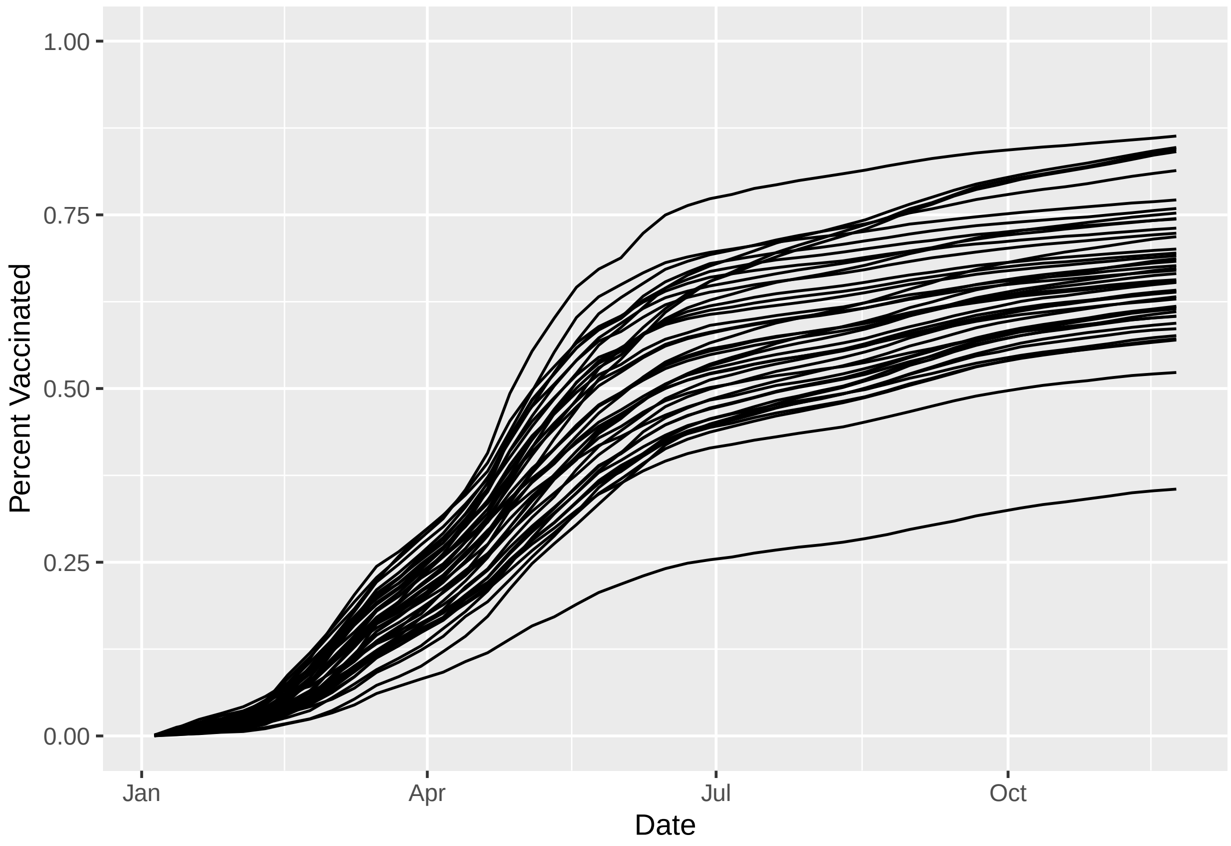
```
length(unique(sd.36$zip_code_tabulation_area))
```

```
## [1] 43
```

Lets make the plot.

```
ggplot(sd.36) +
  aes(x=as_of_date,
      y=percent_of_population_fully_vaccinated,
      group=zip_code_tabulation_area) +
  geom_line() +
  ylim(c(0, 1)) +
  labs(x="Date", y="Percent Vaccinated")
```

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```



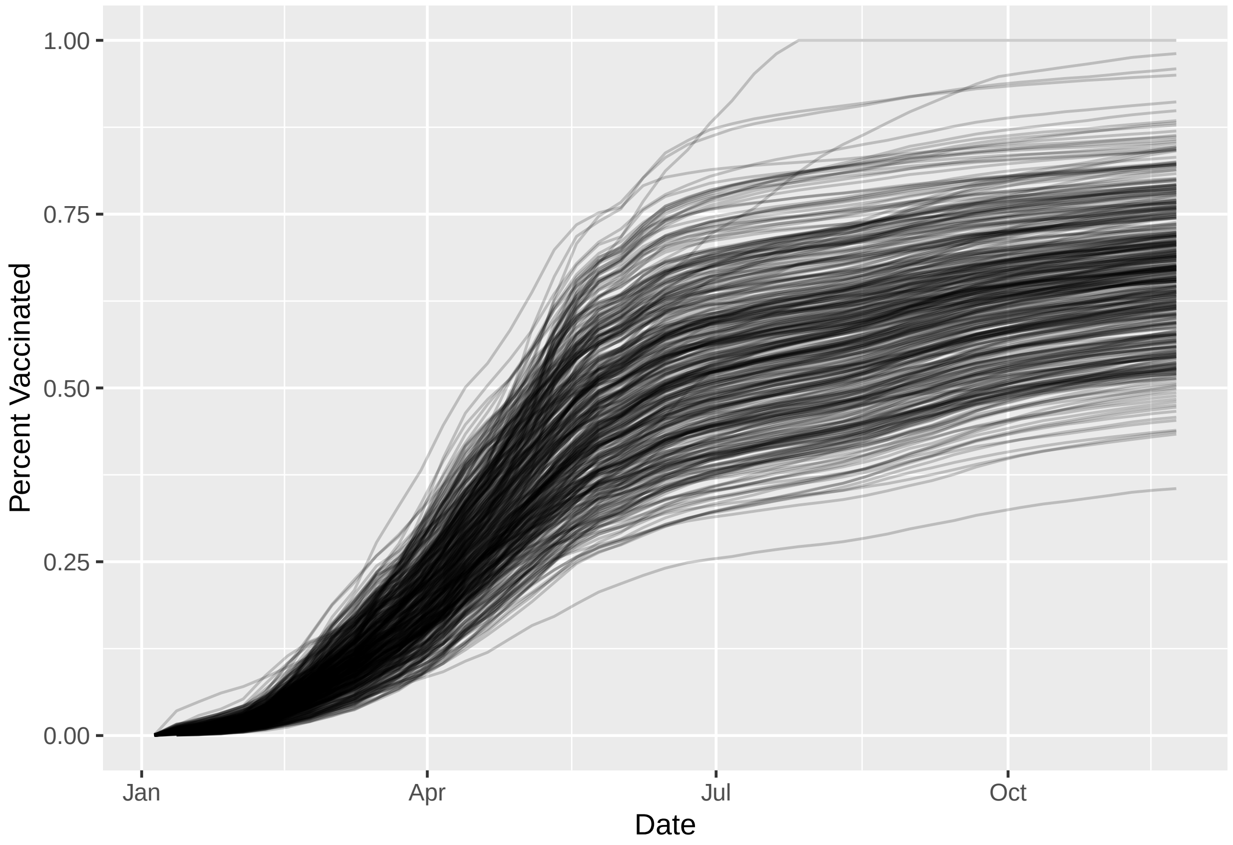
> Q. Make a plot like this for the all ZIP code areas in the State with a population at least as large as La Jolla.

```
ca <- filter(vax, age5_plus_population > 36144)
length(unique(ca$zip_code_tabulation_area))
```

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

```
ggplot(ca) +
  aes(x=as_of_date,
      y=percent_of_population_fully_vaccinated,
      group=zip_code_tabulation_area) +
  geom_line(alpha=0.2) +
  ylim(c(0, 1)) +
  labs(x="Date", y="Percent Vaccinated")
```

```
## Warning: Removed 176 row(s) containing missing values (geom_path).
```



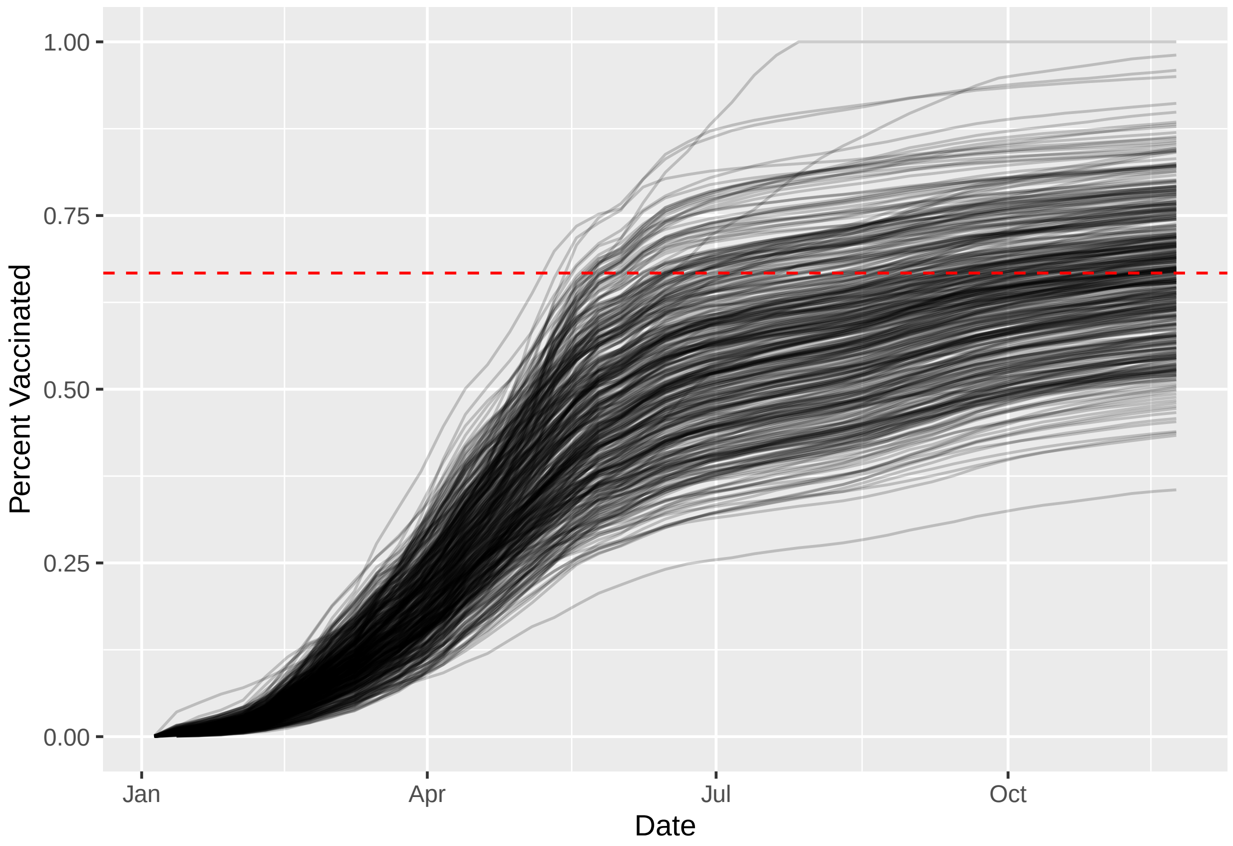
Q. What is the mean across the state for these 36k+ population areas?

```
ca.now <- filter(ca, as_of_date=="2021-11-23")
summary(ca.now$percent_of_population_fully_vaccinated)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.3552  0.5939  0.6696  0.6672  0.7338  1.0000
```

```
ggplot(ca) +
  aes(x=as_of_date,
      y=percent_of_population_fully_vaccinated,
      group=zip_code_tabulation_area) +
  geom_line(alpha=0.2) +
  ylim(c(0, 1)) +
  labs(x="Date", y="Percent Vaccinated") +
  geom_hline(yintercept=mean(ca.now$percent_of_population_fully_vaccinated), linetype=2, color="red")
```

```
## Warning: Removed 176 row(s) containing missing values (geom_path).
```


Percent Vaccinated
Jan
Apr
Jul
Oct
Date