Vaccine Rate Mini Project

Natasha (PID: A15393874)

11/26/2021

Get 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
```

##Ensure the data column is useful

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

```
##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] "2021-11-26"
```

Here we make our 'as_of_date' column lubridate format

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

Now I can do useful math with dates easily:

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

```
## Time difference of 325 days
```

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

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

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

```
## Time difference of 3 days
```

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

```
colnames(vax)
```

```
##  [1] "as_of_date"
##  [2] "zip_code_tabulation_area"
##  [3] "local_health_jurisdiction"
##  [4] "county"
##  [5] "vaccine_equity_metric_quartile"
##  [6] "vem_source"
##  [7] "age12_plus_population"
##  [8] "age5_plus_population"
##  [9] "persons_fully_vaccinated"
## [10] "persons_partially_vaccinated"
## [11] "percent_of_population_fully_vaccinated"
## [12] "percent_of_population_partially_vaccinated"
## [13] "percent_of_population_with_1_plus_dose"
## [14] "redacted"
```

[9] ”persons_fully_vaccinated

Q2. What column details the Zip code tabulation area? [2] “zip_code_tabulation_area”

Q3. What is the earliest date in this dataset?

```
min(vax$as_of_date)
```

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

Q4. What is the latest date in this dataset?

```
max(vax$as_of_date)
```

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

```
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	817.39	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	18895.14	8993.94	0	13468.95	17685.50	21756.25	58556.7	
age5_plus_population	0	1.00	20875.24	9106.04	0	14605.50	19364.00	24877.00	101902.0	
persons_fully_vaccinated	8355	0.90	9585.35	11609.12	21	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? 8355 Q7. What percent of persons_fully_vaccinated values are missing (to 2 significant figures)? 10% of persons_fully_vaccinated values are missing Q8. [Optional]: Why might this data be missing? People not sharing their personal information? Q9. How many days have passed since the last update of the dataset?

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

```
## Time difference of 3 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
```

##Working with zip codes

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

```
#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 San Diego County # Subset to San Diego county only areas

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

```
##    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
```

But let's use the **dplyr** package and it's **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] 5029
```

```
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
```

107 distinct zip codes

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

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

```
##    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
```

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                       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
```

36144

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

```
mean((filter(vax, county == "San Diego" & as_of_date == "2021-11-09"))$percent_of_population_fully_vacci
```

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

67%

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

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

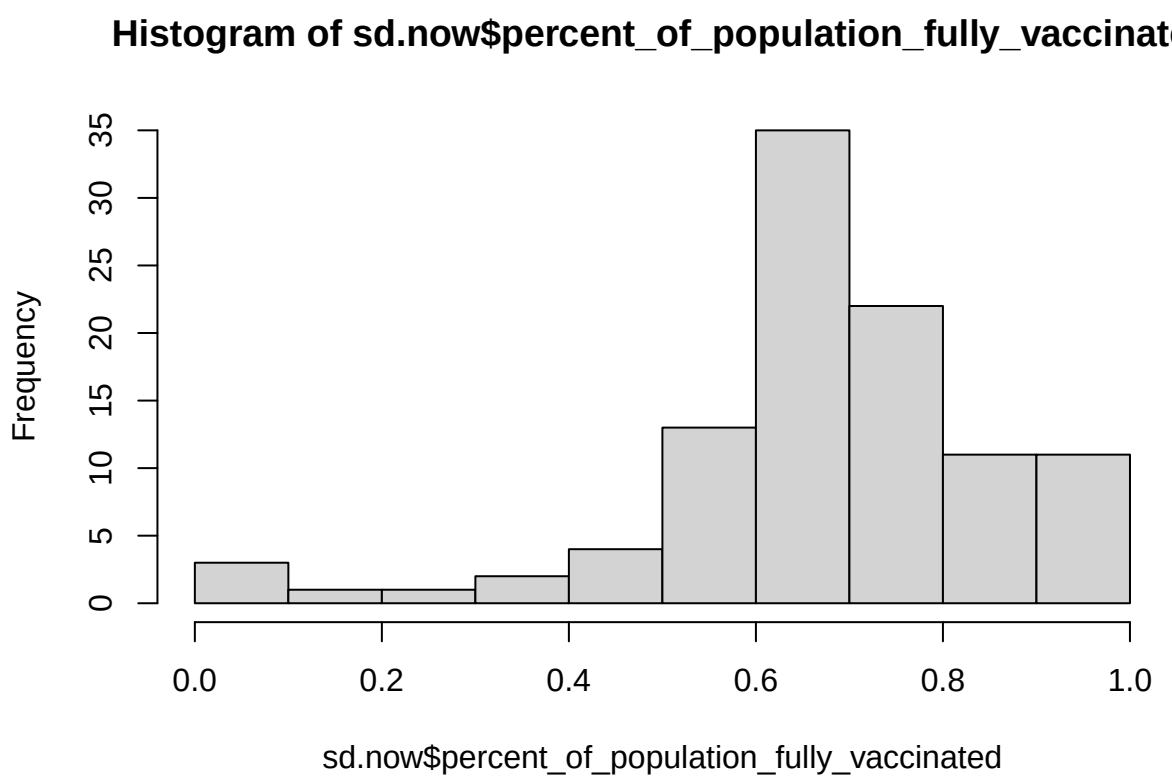
We can look at the 6-number summary

```
sd.sum <- summary(sd.now$percent_of_population_fully_vaccinated)
sd.sum
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.01017 0.60805 0.67711 0.67347 0.76257 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)
```




```
library(ggplot2)
```

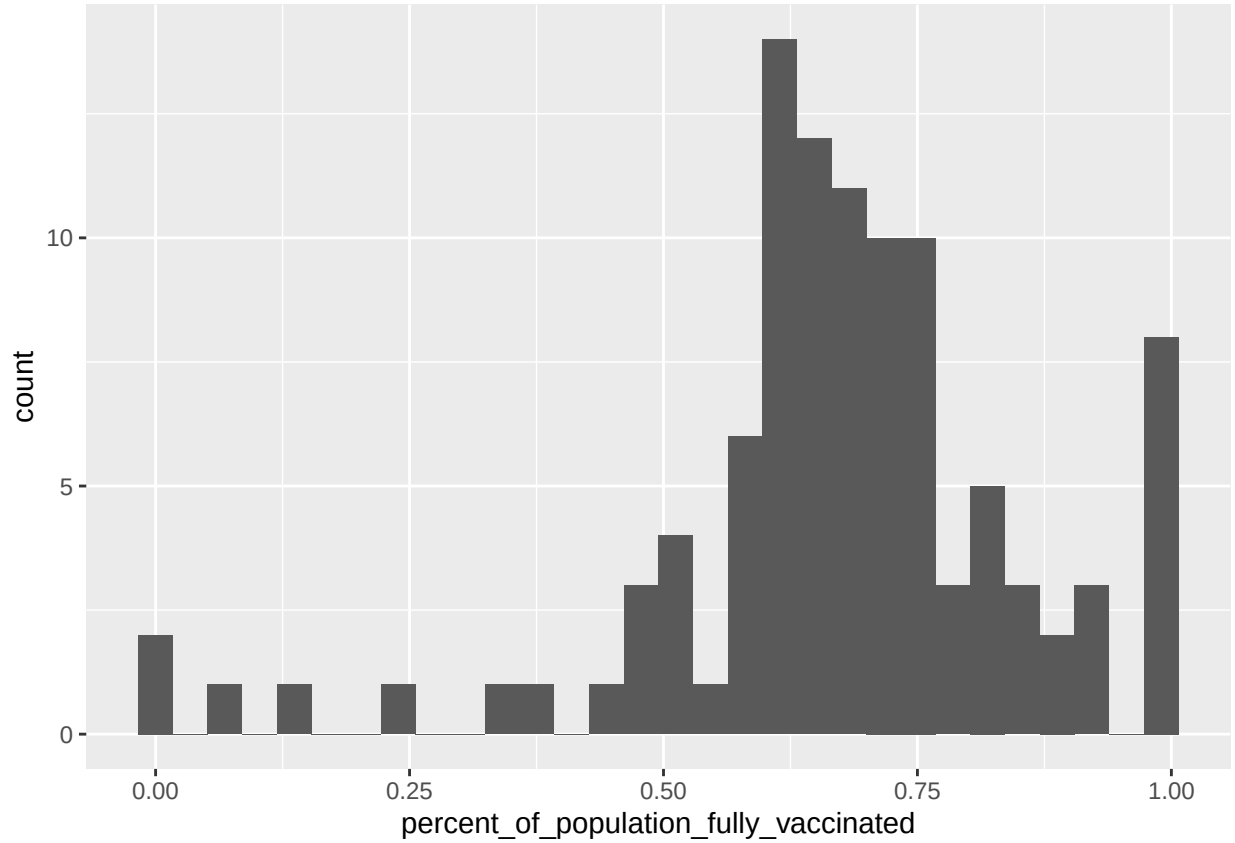
```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

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

```
## Warning: Ignoring unknown parameters: bin
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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



What about 92037 - UCSD/ La Jolla

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

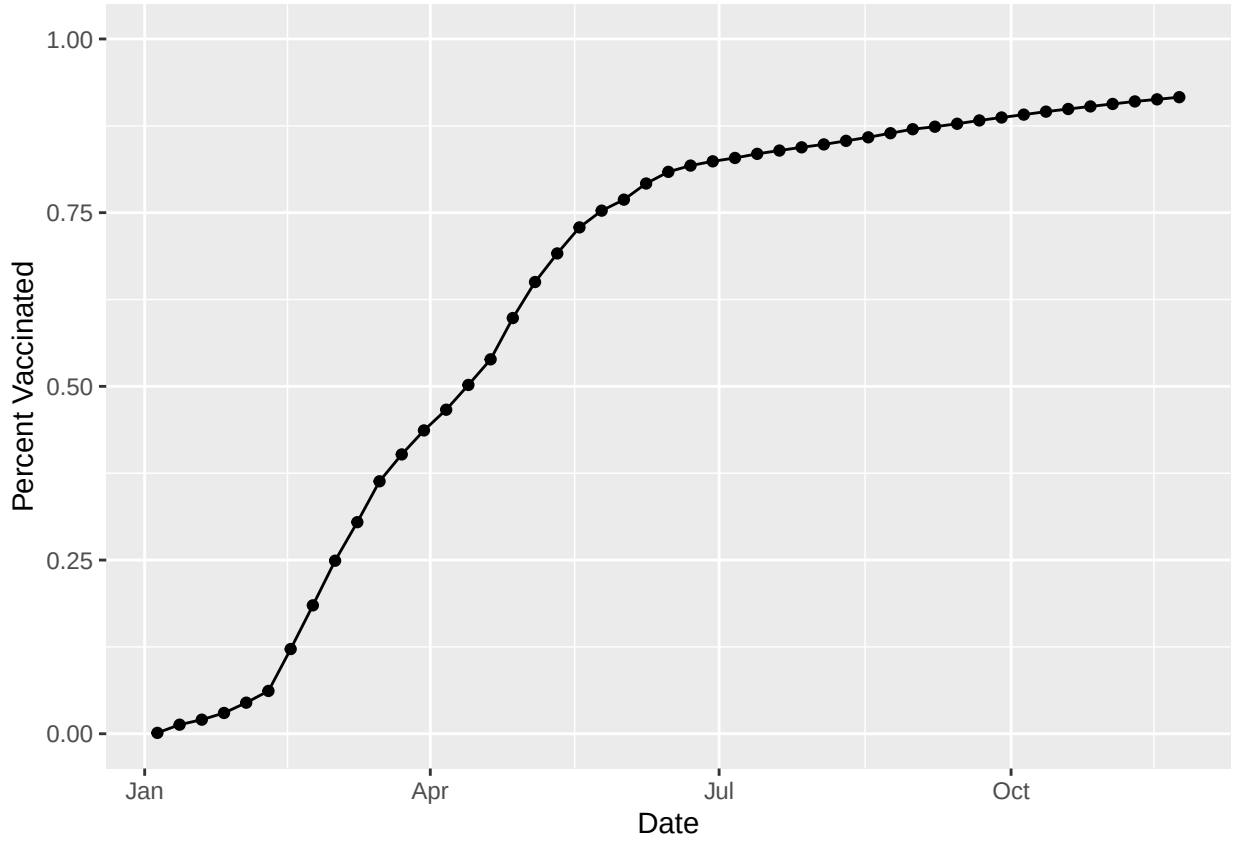
```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 1 2021-11-09                   92037                San Diego San Diego
##   vaccine_equity_metric_quartile               vem_source
## 1                              4 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1               33675.6                36144                    32894
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                         6433                                0.910082
##   percent_of_population_partially_vaccinated
## 1                                   0.177983
##   percent_of_population_with_1_plus_dose redacted
## 1                                      1       No
```

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

```
ucsd <- filter(sd, zip_code_tabulation_area=="92037")
ucsd[1,]$age5_plus_population
```

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

```
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")
```



##Time series of vaccination rate for 92037

```
# 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                   92020                 San Diego      San Diego
## 2 2021-11-16                   92563                 Riverside      Riverside
## 3 2021-11-16                   92806                    Orange         Orange
## 4 2021-11-16                   93291                    Tulare         Tulare
## 5 2021-11-16                   92335           San Bernardino San Bernardino
## 6 2021-11-16                   92618                    Orange         Orange
##   vaccine_equity_metric_quartile                vem_source
## 1                              2 Healthy Places Index Score
## 2                              3 Healthy Places Index Score
```

```
## 3                             2 Healthy Places Index Score
## 4                             1 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               49284.5                54991                     35128
## 2               55897.8                63794                     36051
## 3               33050.9                36739                     24810
## 4               46879.7                54254                     27936
## 5               79670.3                91867                     49820
## 6               40348.0                44304                     39695
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                         5161                                0.638795
## 2                         4224                                0.565116
## 3                         2355                                0.675304
## 4                         4012                                0.514911
## 5                         5970                                0.542306
## 6                         3936                                0.895969
##   percent_of_population_partially_vaccinated
## 1                                   0.093852
## 2                                   0.066213
## 3                                   0.064101
## 4                                   0.073948
## 5                                   0.064985
## 6                                   0.088841
##   percent_of_population_with_1_plus_dose redacted
## 1                               0.732647       No
## 2                               0.631329       No
## 3                               0.739405       No
## 4                               0.588859       No
## 5                               0.607291       No
## 6                               0.984810       No
```

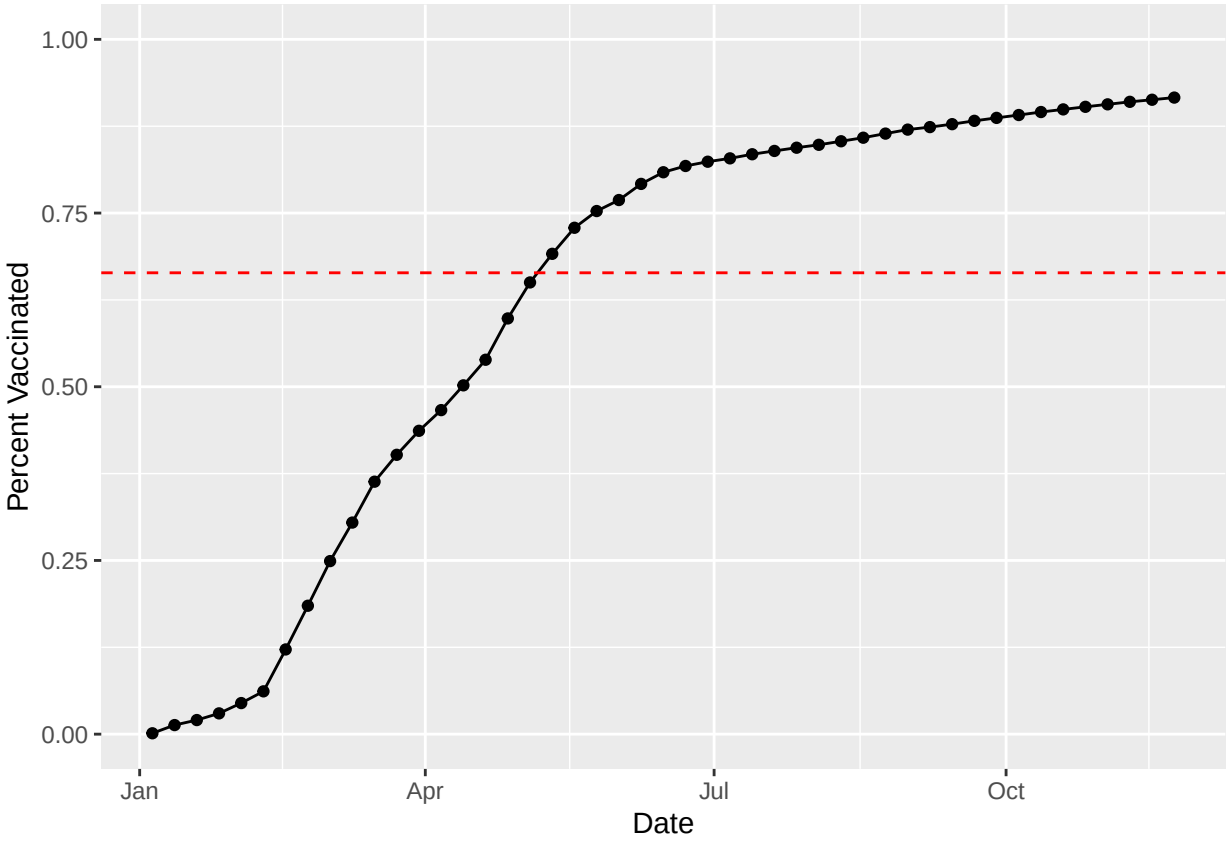
Q16. Calculate the mean "Percent of Population Fully Vaccinated" for ZIP code areas with a population as large as 92037 (La Jolla) as_of_date "2021-11-16". Add this as a straight horizontal line to your plot from above with the geom_hline() function?

```
mean(vax.36$percent_of_population_fully_vaccinated)
```

```
## [1] 0.6640413
```

mean = 66.4%

```
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 = 0.664, colour = "red", linetype = 2)
```



> 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.
##  0.3529  0.5905  0.6662  0.6640  0.7298  1.0000
```

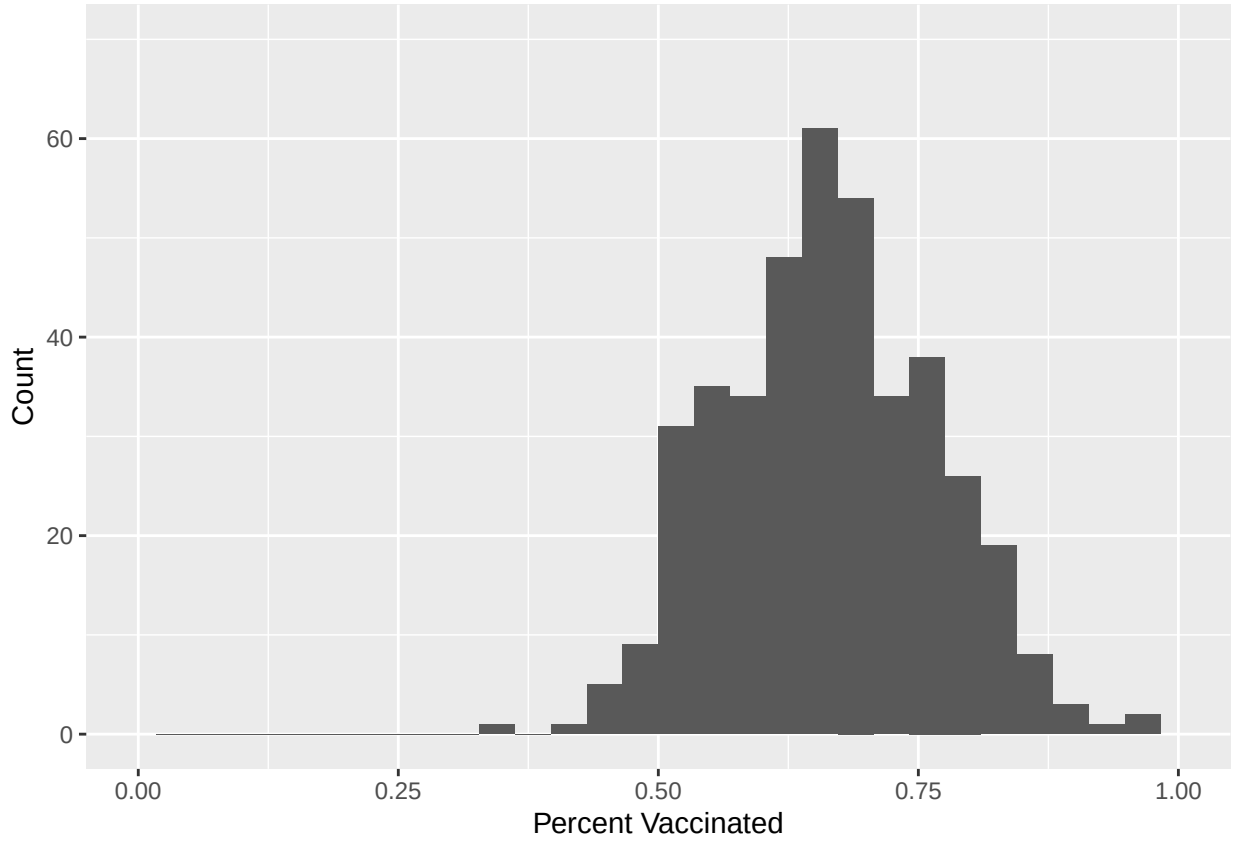
Q18. Using ggplot generate a histogram of this data.

```
ggplot(vax.36) +
  aes(percent_of_population_fully_vaccinated)+geom_histogram(bin=15) +
  labs(x= "Percent Vaccinated", y="Count")+
  xlim(c(0,1))+
  ylim(c(0,70))
```

```
## Warning: Ignoring unknown parameters: bin
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



> Q19. Is the 92109 and 92040 ZIP code areas above or below the average value you calculated for all these above?

```
reverse_zipcode(c('92109', "92040") )
```

```
## # 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 92040   Standard     Lakeside   Lakeside, 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>
```

```
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
```

Lakeside is below the average of 66.4%.

```
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
```

San Diego is above the average of .664.

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

irst we need to subset the full vax dataset to include onl ZIP code areas with a population as large as 92037

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

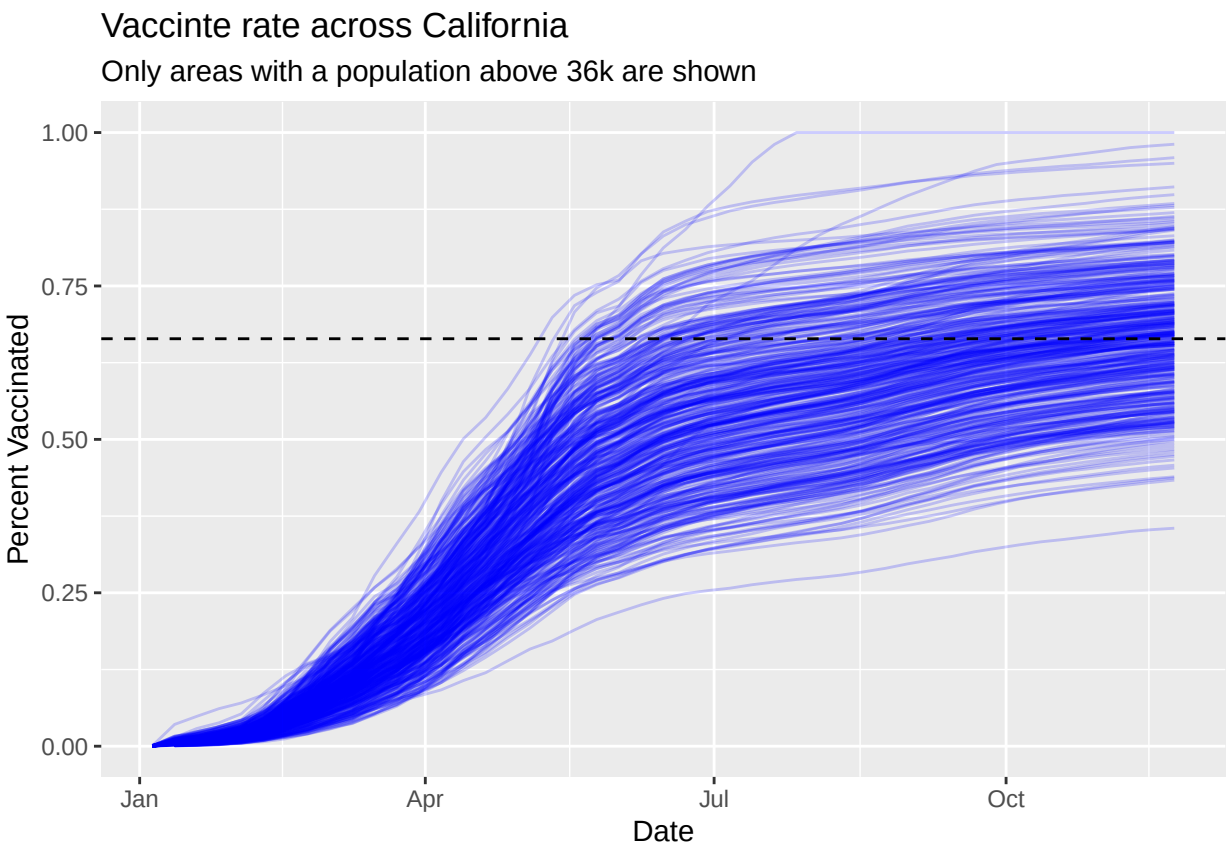
How many unique zip codes have a population as large as 92037?

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

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

```
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(0,1.00)) +
  labs(x="Date", y="Percent Vaccinated",
       title="Vaccinte rate across California",
       subtitle="Only areas with a population above 36k are shown") +
  geom_hline(yintercept = 0.6640, linetype=2)
```

```
## 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? Really nervous because of these statistics and how places are still below average for vaccines and also because of the peak of cases we had last year at this time.