

Class 17 Vaccination Rate Mini Project

Tom Quach (A15549142)

11/23/2021

Download data from CA.gov

```
# 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 persons fully vaccinated column details the total number of people fully vaccinated

Q2. What column details the Zip code tabulation area?

The zip code tabulation area column details the zip code tabulation area

```
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"
```

Q3. What is the earliest date in this dataset?

2021-01-05

```
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"
```

Q4. What is the latest date in this dataset?

2021-11-16

```
##Skirm
```

```
library(skimr)
```

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

Q5. How many numeric columns are in this dataset?

There are 9 numeric columns in this dataset

Q6. Note that there are “missing values” in the dataset. How many NA values there in the `persons_fully_vaccinated` column?

There are 8256 “missing values” for `person_fully_vaccinated` for this dataset

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

```
8256/81144
```

```
## [1] 0.101745
```

10.02% are missing

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

The data is missing may be the result of people are still not fully vaccinated yet

Ensure the date column is useful

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

```
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] "2021-11-28"
```

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

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

```
## Time difference of 327 days
```

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

```
## Time difference of 12 days
```

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

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

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

It has been 7 days between the first and last entry in the dataset

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

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

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

There are a total of 46 unique dates in the dataset

```
## Working with zip codes
```

```
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>
```

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

```
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
```

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

Focus on San Diego findings

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 listed for San Diego County

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
```

The ip code area 92154 in SD has the largest 12+ population in this dataset

What is the population

```
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
```

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")
head(sd.now)
```

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction   county
## 1 2021-11-09                92075                San Diego San Diego
## 2 2021-11-09                92130                San Diego San Diego
## 3 2021-11-09                92060                San Diego San Diego
## 4 2021-11-09                92091                San Diego San Diego
## 5 2021-11-09                92020                San Diego San Diego
## 6 2021-11-09                92004                San Diego San Diego
##   vaccine_equity_metric_quartile                vem_source
## 1                        4 Healthy Places Index Score
## 2                        4 Healthy Places Index Score
## 3                        3 CDPH-Derived ZCTA Score
## 4                        4 CDPH-Derived ZCTA Score
## 5                        2 Healthy Places Index Score
## 6                        2 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1                11136.3                12177                9504
## 2                46300.3                53102                45517
## 3                 166.0                166                153
## 4                 1238.3                1303                1159
## 5                49284.5                54991                34904
## 6                 2151.8                2186                2582
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                        1623                0.780488
## 2                        6642                0.857162
## 3                         34                0.921687
## 4                        221                0.889486
## 5                       4688                0.634722
## 6                        514                1.000000
```

```
## percent_of_population_partially_vaccinated
## 1 0.133284
## 2 0.125080
## 3 0.204819
## 4 0.169609
## 5 0.085250
## 6 0.235133
## percent_of_population_with_1_plus_dose redacted
## 1 0.913772 No
## 2 0.982242 No
## 3 1.000000 No
## 4 1.000000 No
## 5 0.719972 No
## 6 1.000000 No
```

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

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

We can look at the 6-number summary

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

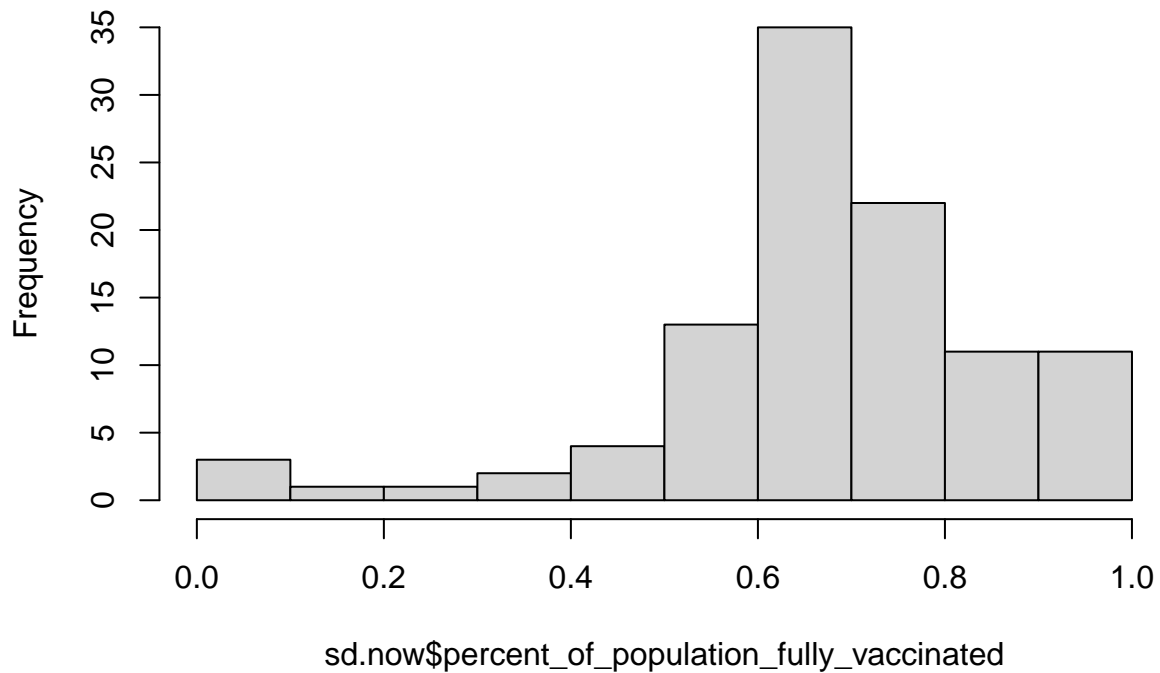
```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NA's
## 0.01017 0.60776 0.67700 0.67276 0.76164 1.00000         4
```

The overall average is 0.6727567

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)
```

Histogram of sd.now\$percent_of_population_fully_vaccinated

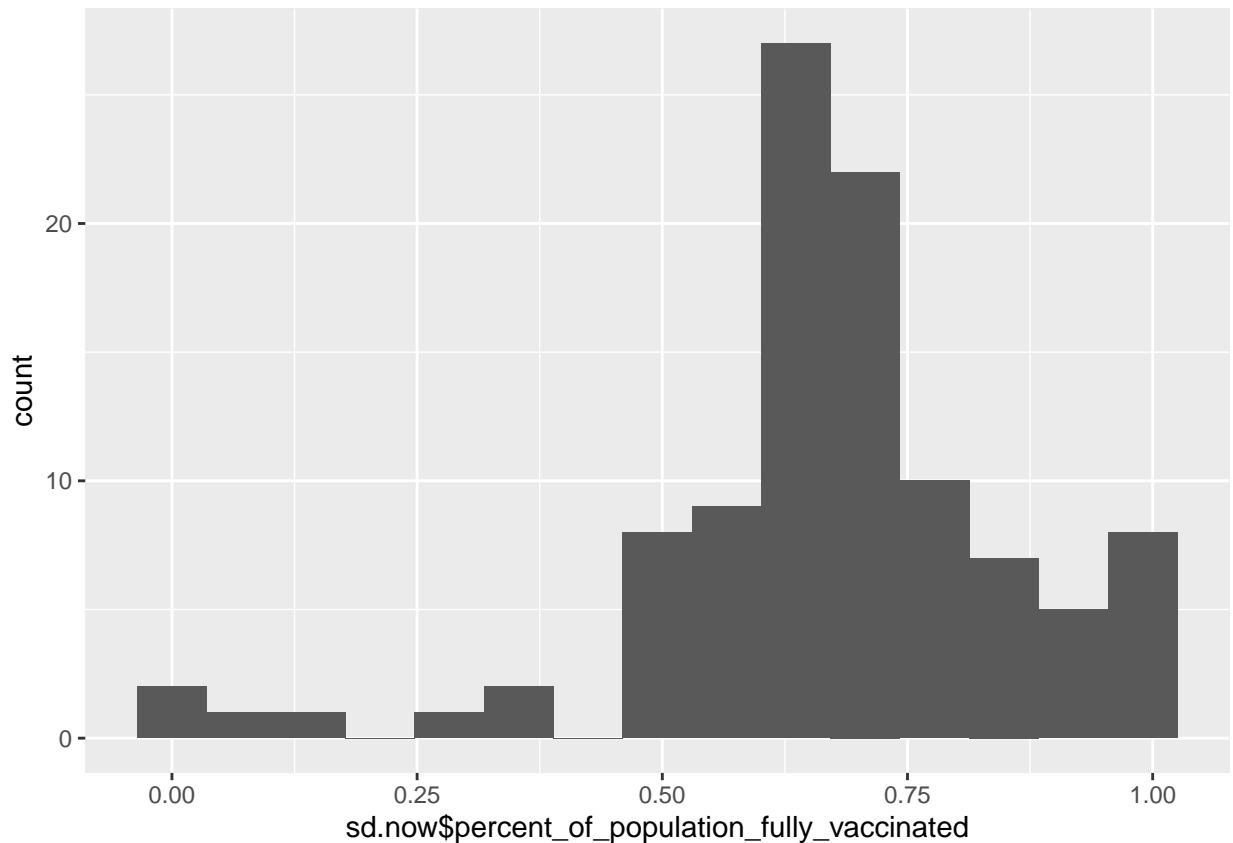


```
library(ggplot2)

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

## Warning: Use of 'sd.now$percent_of_population_fully_vaccinated' is discouraged.
## Use 'percent_of_population_fully_vaccinated' instead.

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

What about 92037 - UCSD/La Jolla

Time series of vaccination rate for 92037

First select all data for the UCSD 92037

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

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

```
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

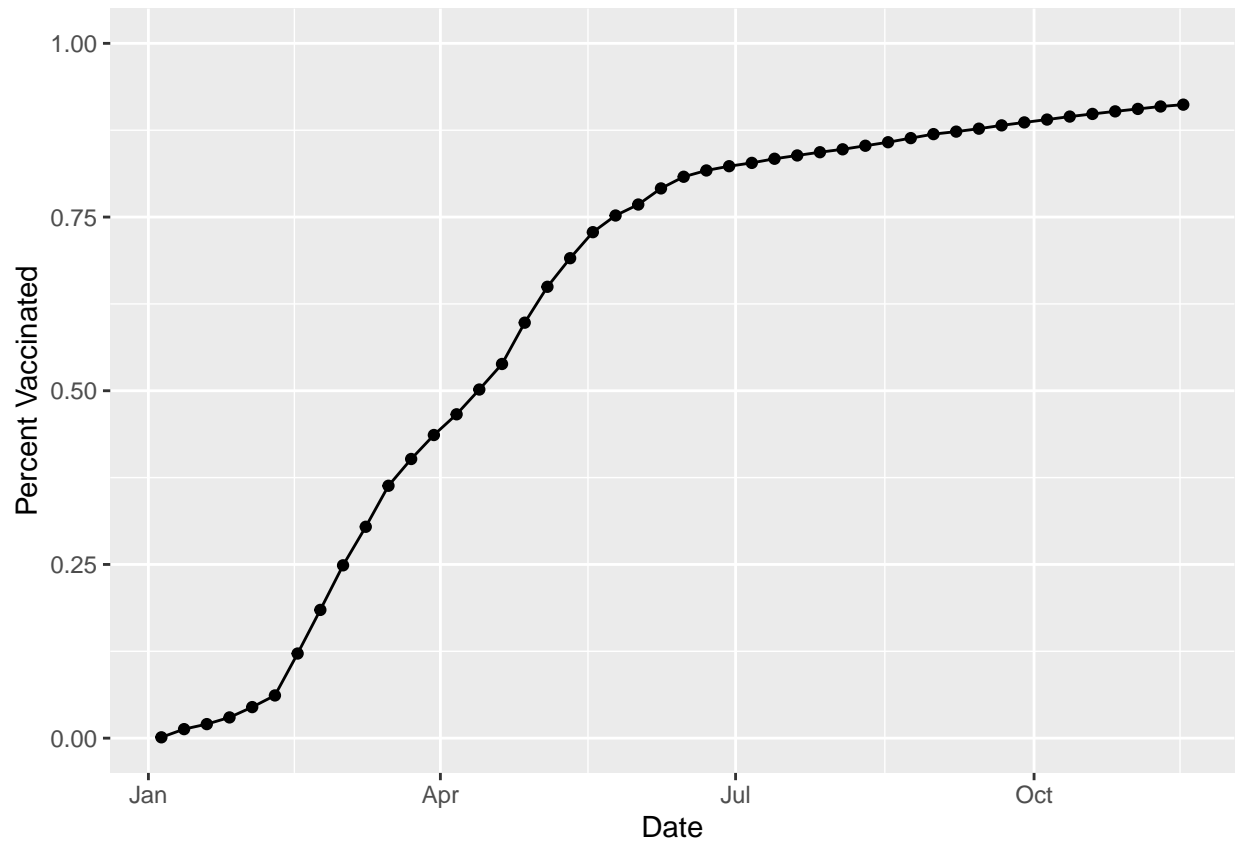
```

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

```

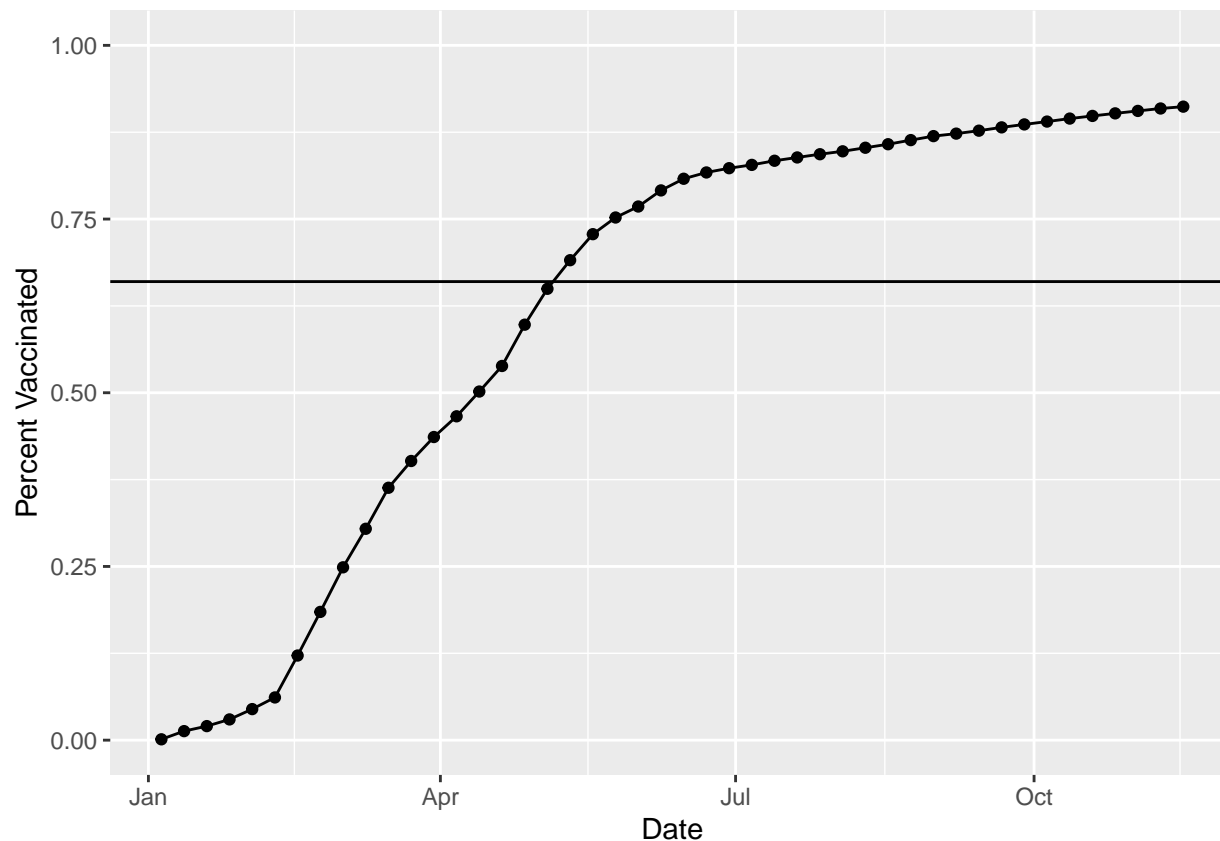
ggplot(ucsd) +
  aes(as_of_date, percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) +
  ylim(c(0,1)) +
  labs(x = "Date", y="Percent Vaccinated")

```



> 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?

```
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)
```



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

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

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”?

```
vax.36 <- filter(vax, age5_plus_population > 36144 & as_of_date == "2021-11-16")
```

```
summary(vax.36)
```

```
##   as_of_date      zip_code_tabulation_area local_health_jurisdiction
## Min.   :2021-11-16  Min.   :90001                Length:411
## 1st Qu.:2021-11-16  1st Qu.:91762                Class :character
## Median :2021-11-16  Median :92646                Mode  :character
## Mean   :2021-11-16  Mean   :92862
## 3rd Qu.:2021-11-16  3rd Qu.:94517
## Max.   :2021-11-16  Max.   :96003
##   county      vaccine_equity_metric_quartile vem_source
## Length:411    Min.   :1.000                Length:411
## Class :character 1st Qu.:1.000                Class :character
## Mode  :character Median :2.000                Mode  :character
##                      Mean   :2.353
```

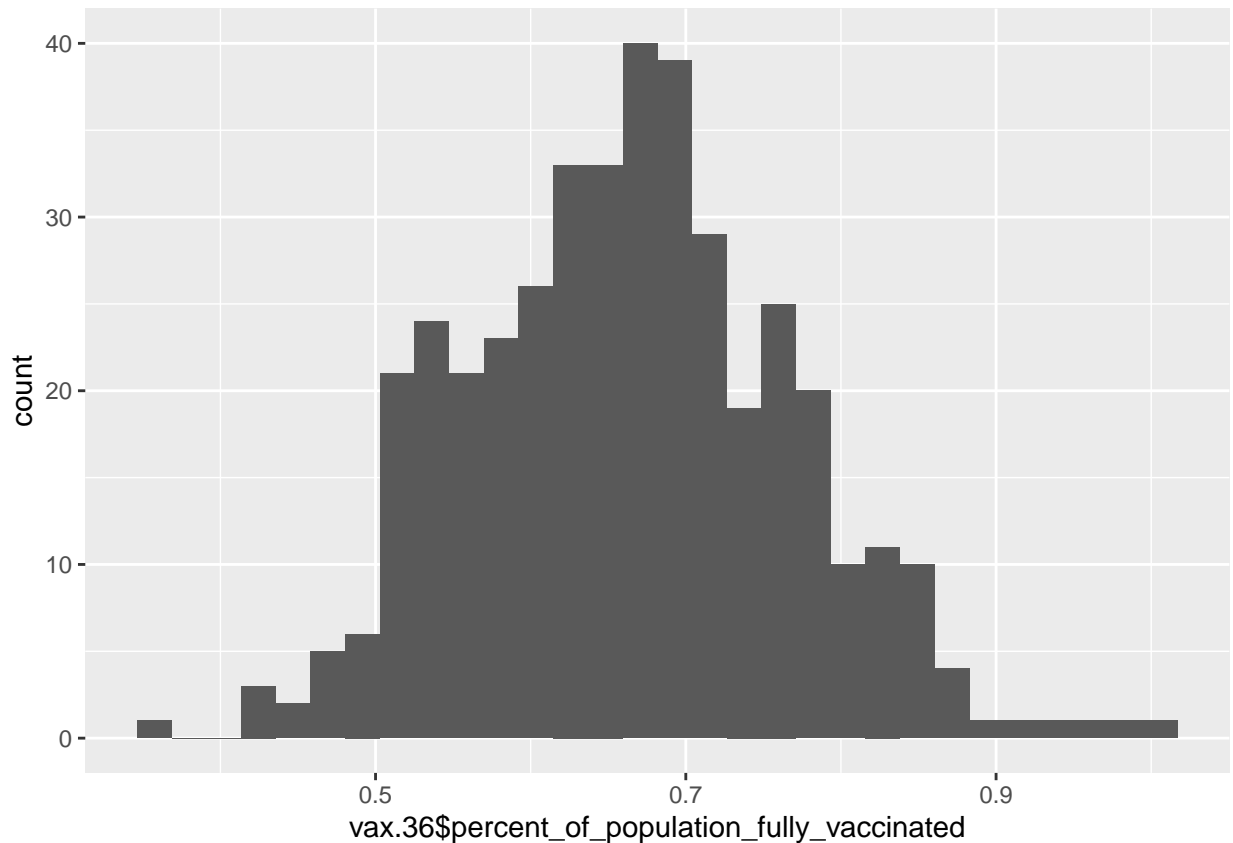
```
##              3rd Qu.:3.000
##              Max.    :4.000
## age12_plus_population age5_plus_population persons_fully_vaccinated
## Min.    :31651      Min.    : 36181      Min.    :13968
## 1st Qu.:37694      1st Qu.: 41613      1st Qu.:27447
## Median :43985      Median : 48573      Median :32322
## Mean   :46847      Mean   : 52012      Mean   :34364
## 3rd Qu.:53932      3rd Qu.: 59168      3rd Qu.:39176
## Max.    :88557      Max.    :101902     Max.    :71078
## persons_partially_vaccinated percent_of_population_fully_vaccinated
## Min.    : 1862      Min.    :0.3519
## 1st Qu.: 2853      1st Qu.:0.5891
## Median : 3532      Median :0.6649
## Mean   : 3917      Mean   :0.6630
## 3rd Qu.: 4524      3rd Qu.:0.7286
## Max.    :14941      Max.    :1.0000
## percent_of_population_partially_vaccinated
## Min.    :0.04658
## 1st Qu.:0.06103
## Median :0.06941
## Mean   :0.07532
## 3rd Qu.:0.08301
## Max.    :0.33035
## percent_of_population_with_1_plus_dose  redacted
## Min.    :0.4168      Length:411
## 1st Qu.:0.6677      Class :character
## Median :0.7377      Mode  :character
## Mean   :0.7371
## 3rd Qu.:0.8068
## Max.    :1.0000
```

Q18. Using ggplot generate a histogram of this data.

```
ggplot(vax.36) + aes(vax.36$percent_of_population_fully_vaccinated) + geom_histogram()
```

```
## Warning: Use of 'vax.36$percent_of_population_fully_vaccinated' is discouraged.
## Use 'percent_of_population_fully_vaccinated' instead.
```

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



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
```

```
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
```

92040 is under and 92109 is above

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

First we need to subset the full “vax” dataset to include only ZIP code areas with a population as large as 92037

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

```
## [1] 18906
```

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

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

Let's make a final figure that shows all these ZIP areas

```
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") +
  geom_hline(yintercept = 0.66, col="red") +
  labs(x="Date", y="Percent of Area Vaccinated")
```

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

