Class 14 Vaccination Rate Mini Project

Cindy Tran

2/13/2022

# 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 94129 San Francisco San Francisco  
## 2 2021-01-05 92562 Riverside Riverside  
## 3 2021-01-05 92805 Orange Orange  
## 4 2021-01-05 92322 San Bernardino San Bernardino  
## 5 2021-01-05 94972 Sonoma Sonoma  
## 6 2021-01-05 94107 San Francisco San Francisco  
## vaccine\_equity\_metric\_quartile vem\_source  
## 1 4 Healthy Places Index Score  
## 2 3 Healthy Places Index Score  
## 3 1 Healthy Places Index Score  
## 4 NA No VEM Assigned  
## 5 NA No VEM Assigned  
## 6 4 Healthy Places Index Score  
## age12\_plus\_population age5\_plus\_population persons\_fully\_vaccinated  
## 1 3574.3 3900 NA  
## 2 53431.1 60184 12  
## 3 61414.4 69071 25  
## 4 581.0 632 NA  
## 5 25.0 25 NA  
## 6 28946.1 30103 12  
## persons\_partially\_vaccinated percent\_of\_population\_fully\_vaccinated  
## 1 NA NA  
## 2 868 0.000199  
## 3 977 0.000362  
## 4 NA NA  
## 5 NA NA  
## 6 836 0.000399  
## percent\_of\_population\_partially\_vaccinated  
## 1 NA  
## 2 0.014422  
## 3 0.014145  
## 4 NA  
## 5 NA  
## 6 0.027771  
## percent\_of\_population\_with\_1\_plus\_dose booster\_recip\_count  
## 1 NA NA  
## 2 0.014621 NA  
## 3 0.014507 NA  
## 4 NA NA  
## 5 NA NA  
## 6 0.028170 NA  
## redacted  
## 1 Information redacted in accordance with CA state privacy requirements  
## 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?

persons\_fully\_vaccinated

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?

2022-02-08

#install.packages("skimr")  
library(skimr)  
skimr::skim(vax)

Data summary

|  |  |
| --- | --- |
| Name | vax |
| Number of rows | 102312 |
| Number of columns | 15 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 5 |
| numeric | 10 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| 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 | 58 | 0 |
| local\_health\_jurisdiction | 0 | 1 | 0 | 15 | 290 | 62 | 0 |
| county | 0 | 1 | 0 | 15 | 290 | 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 | 5046 | 0.95 | 2.44 | 1.11 | 1 | 1.00 | 2.00 | 3.00 | 4.0 | ▇▇▁▇▇ |
| age12\_plus\_population | 0 | 1.00 | 18895.04 | 18993.92 | 0 | 1346.95 | 13685.10 | 31756.12 | 88556.7 | ▇▃▂▁▁ |
| age5\_plus\_population | 0 | 1.00 | 20875.24 | 21106.02 | 0 | 1460.50 | 15364.00 | 34877.00 | 101902.0 | ▇▃▂▁▁ |
| persons\_fully\_vaccinated | 9640 | 0.91 | 10890.58 | 12771.81 | 11 | 623.00 | 5313.00 | 18338.00 | 85970.0 | ▇▂▁▁▁ |
| persons\_partially\_vaccinated | 9640 | 0.91 | 1845.39 | 2062.93 | 11 | 189.00 | 1251.00 | 2790.00 | 29153.0 | ▇▁▁▁▁ |
| percent\_of\_population\_fully\_vaccinated | 9640 | 0.91 | 0.48 | 0.27 | 0 | 0.27 | 0.51 | 0.69 | 1.0 | ▆▅▇▇▃ |
| percent\_of\_population\_partially\_vaccinated | 9640 | 0.91 | 0.09 | 0.11 | 0 | 0.06 | 0.07 | 0.10 | 1.0 | ▇▁▁▁▁ |
| percent\_of\_population\_with\_1\_plus\_dose | 9640 | 0.91 | 0.56 | 0.27 | 0 | 0.37 | 0.59 | 0.76 | 1.0 | ▃▃▆▇▅ |
| booster\_recip\_count | 63642 | 0.38 | 3516.20 | 5246.71 | 11 | 150.00 | 908.00 | 5069.75 | 48283.0 | ▇▁▁▁▁ |

Q5. How many numeric columns are in this dataset?

15

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] 9640

9640

Q7. What percent of persons\_fully\_vaccinated values are missing (to 2 significant figures)?

9640 / (sum(!is.na(vax$persons\_fully\_vaccinated))) \* 100

## [1] 10.40228

10%

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

This data is posisbly missing because some counties did not have collect this information.

## Working with Dates

library(lubridate)

##   
## Attaching package: 'lubridate'

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

today()

## [1] "2022-02-14"

# Specify that we are using the Year-month-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 405 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 399 days

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

today() - vax$as\_of\_date[nrow(vax)]

## Time difference of 6 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] 58

# Working with ZIP Codes

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

Find the centroid of the La Jolla 92037 (i.e. UC San Diego) ZIP code area.

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

zip\_distance('92037','92109')

## zipcode\_a zipcode\_b distance  
## 1 92037 92109 2.33

We can 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>

# Pull data for all ZIP codes in the dataset  
zipdata <- reverse\_zipcode( vax$zip\_code\_tabulation\_area )

# Focus on the San Diego Area

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

## [1] 6206

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] 6206

Using dplyr is often more convenient when we are subsetting across multiple criteria - 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

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

which.max(sd$age12\_plus\_population)

## [1] 56

sd[56,]

## as\_of\_date zip\_code\_tabulation\_area local\_health\_jurisdiction county  
## 56 2021-01-05 92154 San Diego San Diego  
## vaccine\_equity\_metric\_quartile vem\_source  
## 56 2 Healthy Places Index Score  
## age12\_plus\_population age5\_plus\_population persons\_fully\_vaccinated  
## 56 76365.2 82971 33  
## persons\_partially\_vaccinated percent\_of\_population\_fully\_vaccinated  
## 56 1357 0.000398  
## percent\_of\_population\_partially\_vaccinated  
## 56 0.016355  
## percent\_of\_population\_with\_1\_plus\_dose booster\_recip\_count  
## 56 0.016753 NA  
## redacted  
## 56 Information redacted in accordance with CA state privacy requirements

92154

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

q13 <- filter (sd, as\_of\_date == "2021-11-09")  
mean(q13$percent\_of\_population\_fully\_vaccinated, na.rm = TRUE)

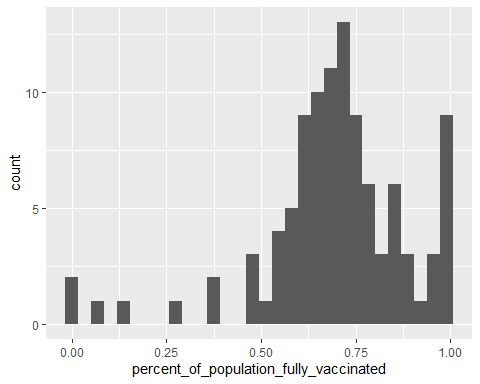
## [1] 0.6961169

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”?

library(ggplot2)  
ggplot(q13) +  
 aes(x = percent\_of\_population\_fully\_vaccinated) +   
 geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 4 rows containing non-finite values (stat\_bin).



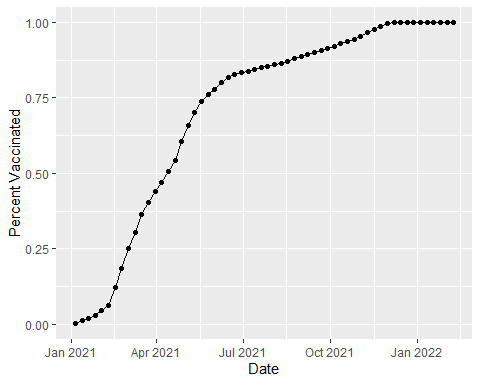
## Focus on UCSD/La Jolla

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:

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 SImilarly 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)

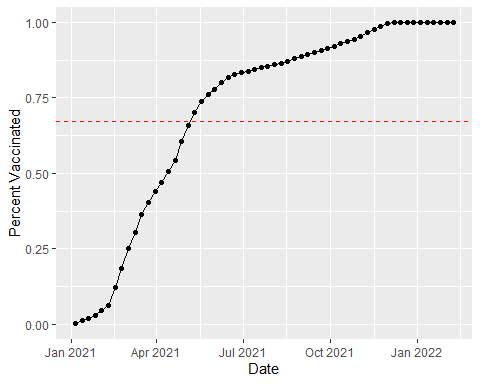
## as\_of\_date zip\_code\_tabulation\_area local\_health\_jurisdiction county  
## 1 2021-11-16 93063 Ventura Ventura  
## 2 2021-11-16 92591 Riverside Riverside  
## 3 2021-11-16 91745 Los Angeles Los Angeles  
## 4 2021-11-16 93311 Kern Kern  
## 5 2021-11-16 95240 San Joaquin San Joaquin  
## 6 2021-11-16 92505 Riverside Riverside  
## vaccine\_equity\_metric\_quartile vem\_source  
## 1 4 Healthy Places Index Score  
## 2 3 Healthy Places Index Score  
## 3 3 Healthy Places Index Score  
## 4 3 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 49342.3 53192 35688  
## 2 34147.8 38439 21584  
## 3 48344.2 52318 39646  
## 4 37656.8 42439 30104  
## 5 39228.8 44646 24225  
## 6 44919.3 50178 27181  
## persons\_partially\_vaccinated percent\_of\_population\_fully\_vaccinated  
## 1 2933 0.670928  
## 2 2516 0.561513  
## 3 3265 0.757789  
## 4 3286 0.709348  
## 5 4228 0.542602  
## 6 2947 0.541692  
## percent\_of\_population\_partially\_vaccinated  
## 1 0.055140  
## 2 0.065454  
## 3 0.062407  
## 4 0.077429  
## 5 0.094701  
## 6 0.058731  
## percent\_of\_population\_with\_1\_plus\_dose booster\_recip\_count redacted  
## 1 0.726068 7001 No  
## 2 0.626967 3487 No  
## 3 0.820196 8195 No  
## 4 0.786777 5635 No  
## 5 0.637303 3069 No  
## 6 0.600423 3271 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, na.rm = TRUE)

## [1] 0.6716873

ggplot(ucsd) +  
 aes(as\_of\_date, percent\_of\_population\_fully\_vaccinated) +   
 geom\_hline(yintercept = 0.6716873, linetype = "dashed", color = "red") +  
 geom\_point() +  
 geom\_line(group=1) +  
 ylim(c(0,1)) +  
 labs(x="Date", y="Percent Vaccinated")



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”?

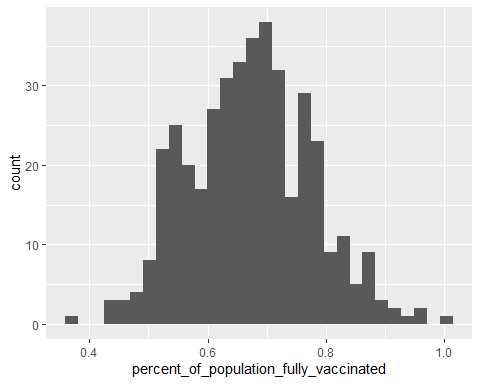
summary(vax.36$percent\_of\_population\_fully\_vaccinated)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.3675 0.5992 0.6738 0.6717 0.7408 1.0000

Q18. Using ggplot generate a histogram of this data.

ggplot(vax.36) +  
 aes(x = percent\_of\_population\_fully\_vaccinated) +  
 geom\_histogram()

## `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=="92109") %>%  
 select(percent\_of\_population\_fully\_vaccinated)

## percent\_of\_population\_fully\_vaccinated  
## 1 0.717349

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 0.535312

92109 is above the average while 92040 is below the average.

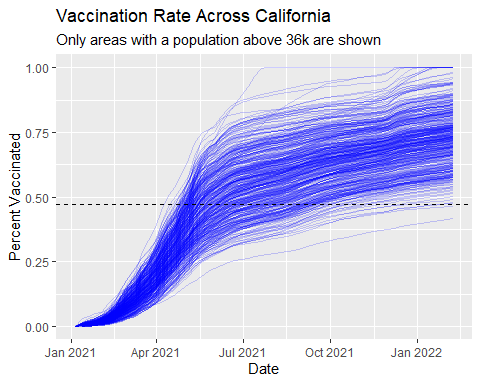
Q20. Finally make a time course plot of vaccination progress for all areas in the full dataset with a age5\_plus\_population > 36144

vax.36.all <- filter(vax, age5\_plus\_population > 36144)  
  
mean(vax.36.all$percent\_of\_population\_fully\_vaccinated, na.rm = TRUE)

## [1] 0.472364

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(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 = 0.472364, linetype="dashed")

## Warning: Removed 174 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 uncomfortable going to in person classes, but am okay with going if necessary.

sessionInfo()

## R version 4.1.2 (2021-11-01)  
## Platform: x86\_64-w64-mingw32/x64 (64-bit)  
## Running under: Windows 10 x64 (build 19043)  
##   
## Matrix products: default  
##   
## locale:  
## [1] LC\_COLLATE=English\_United States.1252   
## [2] LC\_CTYPE=English\_United States.1252   
## [3] LC\_MONETARY=English\_United States.1252  
## [4] LC\_NUMERIC=C   
## [5] LC\_TIME=English\_United States.1252   
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## other attached packages:  
## [1] ggplot2\_3.3.5 dplyr\_1.0.7 zipcodeR\_0.3.3 lubridate\_1.8.0  
## [5] skimr\_2.1.3   
##   
## loaded via a namespace (and not attached):  
## [1] httr\_1.4.2 tidyr\_1.2.0 bit64\_4.0.5 jsonlite\_1.7.3   
## [5] assertthat\_0.2.1 sp\_1.4-6 highr\_0.9 blob\_1.2.2   
## [9] yaml\_2.2.1 tidycensus\_1.1 pillar\_1.7.0 RSQLite\_2.2.9   
## [13] lattice\_0.20-45 glue\_1.6.0 uuid\_1.0-3 digest\_0.6.27   
## [17] rvest\_1.0.2 colorspace\_2.0-2 htmltools\_0.5.1.1 pkgconfig\_2.0.3   
## [21] raster\_3.5-15 purrr\_0.3.4 scales\_1.1.1 terra\_1.5-17   
## [25] tzdb\_0.2.0 tigris\_1.5.1 tibble\_3.1.6 proxy\_0.4-26   
## [29] farver\_2.1.0 generics\_0.1.2 ellipsis\_0.3.2 withr\_2.4.3   
## [33] cachem\_1.0.6 repr\_1.1.4 cli\_3.1.1 magrittr\_2.0.1   
## [37] crayon\_1.4.2 memoise\_2.0.1 maptools\_1.1-2 evaluate\_0.14   
## [41] fansi\_1.0.2 xml2\_1.3.3 foreign\_0.8-82 class\_7.3-20   
## [45] tools\_4.1.2 hms\_1.1.1 lifecycle\_1.0.1 stringr\_1.4.0   
## [49] munsell\_0.5.0 compiler\_4.1.2 e1071\_1.7-9 rlang\_0.4.11   
## [53] classInt\_0.4-3 units\_0.8-0 grid\_4.1.2 rstudioapi\_0.13   
## [57] rappdirs\_0.3.3 labeling\_0.4.2 base64enc\_0.1-3 rmarkdown\_2.11   
## [61] gtable\_0.3.0 codetools\_0.2-18 DBI\_1.1.2 curl\_4.3.2   
## [65] R6\_2.5.1 knitr\_1.37 rgdal\_1.5-28 fastmap\_1.1.0   
## [69] bit\_4.0.4 utf8\_1.2.2 KernSmooth\_2.23-20 readr\_2.1.2   
## [73] stringi\_1.7.6 Rcpp\_1.0.8 vctrs\_0.3.8 sf\_1.0-6   
## [77] tidyselect\_1.1.1 xfun\_0.29