# BIMM 143 Lab 17 Vaccination Rates

Patric Young

## COVID-19 Vaccination Rates

```
vax <- read.csv("covid19vaccinesbyzipcode_test.csv")
head(vax)
```

```
  as_of_date zip_code_tabulation_area local_health_jurisdiction    county
1 2021-01-05                   95446                   Sonoma    Sonoma
2 2021-01-05                   96014                  Siskiyou  Siskiyou
3 2021-01-05                   96087                   Shasta    Shasta
4 2021-01-05                   96008                   Shasta    Shasta
5 2021-01-05                   95410                 Mendocino Mendocino
6 2021-01-05                   95527                   Trinity   Trinity
  vaccine_equity_metric_quartile                vem_source
1                              2 Healthy Places Index Score
2                              2    CDPH-Derived ZCTA Score
3                              2    CDPH-Derived ZCTA Score
4                             NA            No VEM Assigned
5                              3    CDPH-Derived ZCTA Score
6                              2    CDPH-Derived ZCTA Score
  age12_plus_population age5_plus_population tot_population
1                4840.7                 5057           5168
2                 135.0                  135            135
3                 513.9                  544            544
4                1125.3                 1164             NA
5                 926.3                  988            997
6                 476.6                  485            499
  persons_fully_vaccinated persons_partially_vaccinated
1                       NA                           NA
2                       NA                           NA
3                       NA                           NA
```

```
4                       NA                            NA
5                       NA                            NA
6                       NA                            NA
  percent_of_population_fully_vaccinated
1                                     NA
2                                     NA
3                                     NA
4                                     NA
5                                     NA
6                                     NA
  percent_of_population_partially_vaccinated
1                                         NA
2                                         NA
3                                         NA
4                                         NA
5                                         NA
6                                         NA
  percent_of_population_with_1_plus_dose booster_recip_count
1                                     NA                  NA
2                                     NA                  NA
3                                     NA                  NA
4                                     NA                  NA
5                                     NA                  NA
6                                     NA                  NA
  bivalent_dose_recip_count eligible_recipient_count
1                        NA                       0
2                        NA                       0
3                        NA                       2
4                        NA                       2
5                        NA                       0
6                        NA                       0
                                                                    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?

```
#vax$persons_fully_vaccinated
```

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

```
#vax$zip_code_tabulation_area
```

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

```
vax$as_of_date[1]
```

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

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

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

```
[1] "2023-02-28"
```

We can use the skim() function for a quick overview of a new data set like this.

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

Table 1: Data summary

| | |
|---|---|
| Name | vax |
| Number of rows | 199332 |
| Number of columns | 18 |
| _____ | |
| Column type frequency: | |
| character | 5 |
| numeric | 13 |
| _____ | |
| 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 | 113 | 0 |
| local_health_jurisdiction | 0 | 1 | 0 | 15 | 565 | 62 | 0 |
| county | 0 | 1 | 0 | 15 | 565 | 59 | 0 |

| skim_variable | n_missing | complete_rate | min | max | empty | n_unique | whitespace |
|---|---|---|---|---|---|---|---|
| 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 | 1817.38 | 90001 | 92257.75 | 93658.50 | 95380.50 | 97635.0 | |
| vaccine_equity_metric_quartile | 9831 | 0.95 | 2.44 | 1.11 | 1 | 1.00 | 2.00 | 3.00 | 4.0 | |
| age12_plus_population | 0 | 1.00 | 18895.01 | 8993.87 | 0 | 13346.95 | 13685.50 | 31756.00 | 58556.7 | |
| age5_plus_population | 0 | 1.00 | 20875.22 | 11105.97 | 0 | 14605.50 | 15364.00 | 34877.00 | 101902.0 | |
| tot_population | 9718 | 0.95 | 23372.72 | 12628.51 | 12 | 21260.00 | 18714.00 | 38168.00 | 111165.0 | |
| persons_fully_vaccinated | 16525 | 0.92 | 13962.33 | 5054.09 | 11 | 9300.00 | 8566.00 | 23302.00 | 37566.0 | |
| persons_partially_vaccinated | 16525 | 0.92 | 1701.64 | 2030.18 | 11 | 1650.00 | 1196.00 | 2535.00 | 39913.0 | |
| percent_of_population_fully_vaccinated | 20825 | 0.90 | 0.57 | 0.25 | 0 | 0.42 | 0.60 | 0.74 | 1.0 | |
| percent_of_population_partially_vaccinated | 20825 | 0.90 | 0.08 | 0.09 | 0 | 0.05 | 0.06 | 0.08 | 1.0 | |
| percent_of_population_with_1_plus_dose | 21859 | 0.89 | 0.63 | 0.24 | 0 | 0.49 | 0.67 | 0.81 | 1.0 | |
| booster_recip_count | 72872 | 0.63 | 5837.31 | 7165.81 | 11 | 2970.00 | 2748.00 | 9438.25 | 59553.0 | |
| bivalent_dose_recip_count | 158664 | 0.20 | 2924.93 | 3583.45 | 11 | 1900.00 | 1418.00 | 4626.25 | 27458.0 | |
| eligible_recipient_count | 0 | 1.00 | 12801.84 | 4908.33 | 0 | 5040.00 | 6338.00 | 21973.00 | 37234.0 | |

**Q5.** How many numeric columns are in this data set?

13

**Q6.** Note that there are "missing values" in the data set. How many `NA` values there in the `persons_fully_vaccinated` column?

```
n.missing <- sum(is.na(vax$persons_fully_vaccinated))
n.missing
```

```
[1] 16525
```

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

```
round((n.missing/ nrow(vax))*100,2)
```

```
[1] 8.29
```

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

### Working with dates

The `lubridate` package makes working with dates and times in R much less of a pain. Let's have a first play with this package here.

```
library(lubridate)
```

```
Attaching package: 'lubridate'

The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

```
today()
```

```
[1] "2023-03-07"
```

We can now magically do math with dates.

```
today() - ymd("2021-01-05")
```

```
Time difference of 791 days
```

How old am I?

```
today() - ymd("2001-03-15")
```

```
Time difference of 8027 days
```

Let's treat the whole col.

How many days have passed since the first vaccination reported in this data set.

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

```
Time difference of 791 days
```

How many days does the dataset span.

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

```
Time difference of 784 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
```

**Q10.** How many unique dates are in the dataset (i.e. how many different dates are detailed)?67

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

```
[1] 113
```

```
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
```

```
n_distinct(vax$as_of_date)
```

```
[1] 113
```

## Zip Codes

ZIP codes are rather annoying things to work with as they are numeric but not in the conventional sense of doing math.

Just like dates we have special packages to help us work with

```
library(zipcodeR)
```

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

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

```
reverse_zipcode(c('92037', "92109") )
```

```
# A tibble: 2 x 24
  zipcode zipcode_~1 major~2 post_~3 common_c~4 county state   lat   lng timez~5
  <chr>   <chr>      <chr>   <chr>       <blob> <chr>  <chr> <dbl> <dbl> <chr>
1 92037   Standard   La Jol~ La Jol~ <raw 20 B> San D~ CA     32.8 -117. Pacific
2 92109   Standard   San Di~ San Di~ <raw 21 B> San D~ CA     32.8 -117. Pacific
# ... with 14 more variables: 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>, and abbreviated variable names
#   1: zipcode_type, 2: major_city, 3: post_office_city, ...
```

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

```
[1] 0
```

```
library(dplyr)

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

nrow(sd)
```

```
[1] 12091
```

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

```
[1] 8588
```

```
n_distinct(sd.10$zip_code_tabulation_area)
```

```
[1] 76
```

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

```
n_distinct(sd$zip_code_tabulation_area)
```

```
[1] 107
```

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

```
ind <- which.max(sd$age12_plus_population)
sd$zip_code_tabulation_area [2]
```

```
[1] 92154
```

**Q13.** What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2023-11-15"?

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

```
[1] "2023-02-28"
```

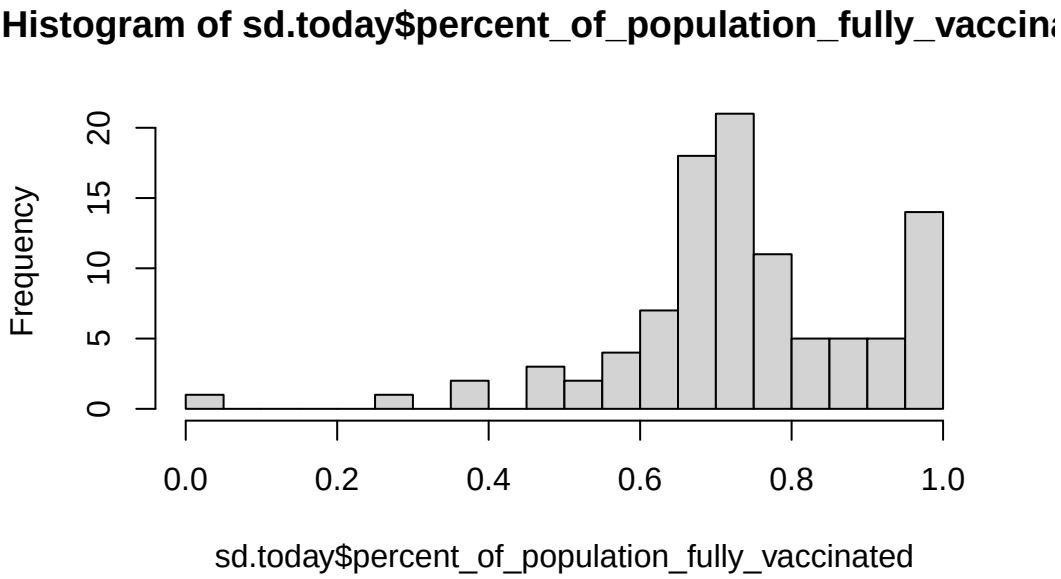
```
sd.today <- filter(sd, as_of_date == "2023-02-28")
```

```
mean(sd.today$percent_of_population_fully_vaccinated, na.rm=T)
```

```
[1] 0.7400878
```

**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 "2022-11-15"?

```
hist(sd.today$percent_of_population_fully_vaccinated, breaks=20)
```



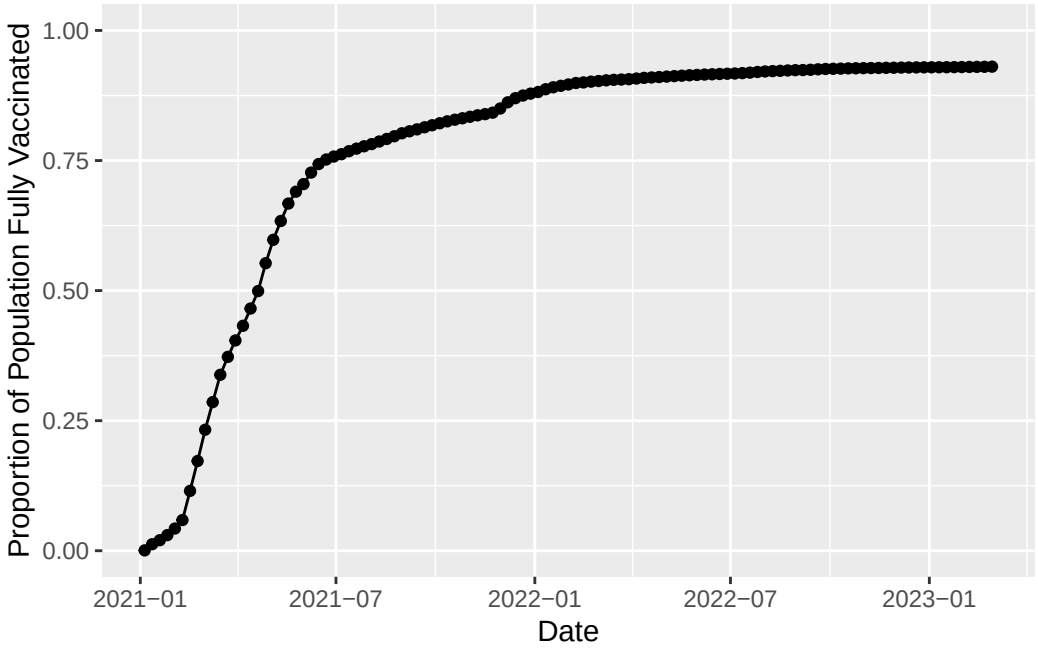

### Focus on UCSD/La Jolla

```
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:

```
library(ggplot2)
ucplot <- ggplot(ucsd) +
  aes(as_of_date, percent_of_population_fully_vaccinated) +
  geom_point()+
  geom_line(group=1) +
  xlab("Date")+
ylab("Proportion of Population Fully Vaccinated")+
  ylim(c(0,1))
ucplot
```



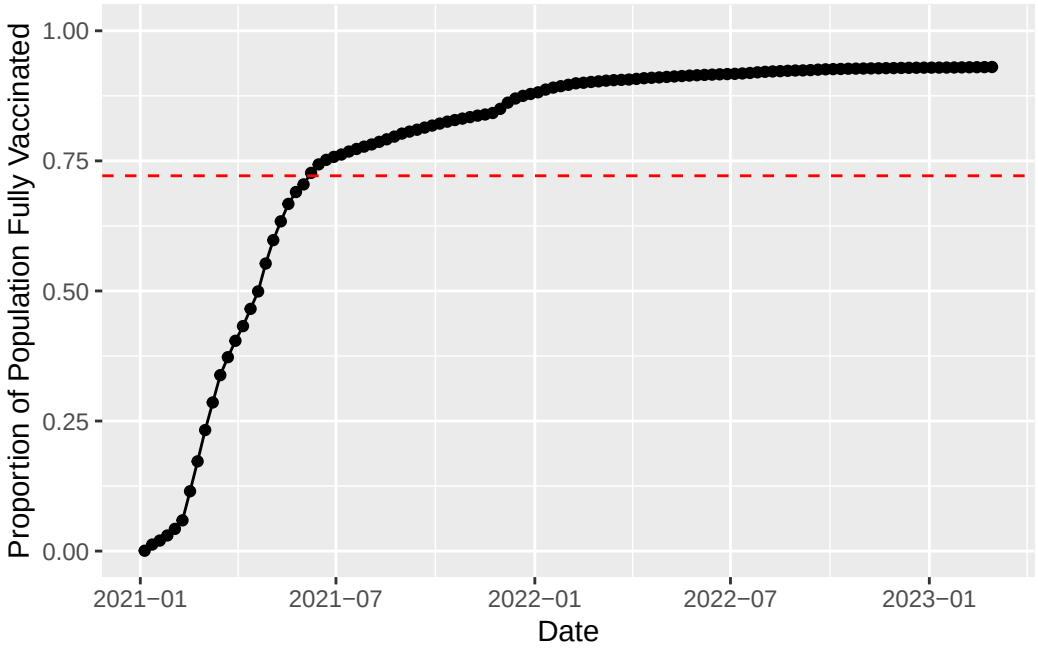

```
# Subset to all CA areas with a population as large as 92037
vax.36 <- filter(vax, age5_plus_population > 36144 &
                 as_of_date == "2023-02-28")

#head(vax.36)
```

**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* "2022-11-15". Add this as a straight horizontal line to your plot from above with the `geom_hline()` function?

```
ave <- mean(vax.36$percent_of_population_fully_vaccinated)
ave
```

```
[1] 0.7213331
```

```
ucplot + geom_hline(yintercept = ave, col="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* "2022-11-15"?

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

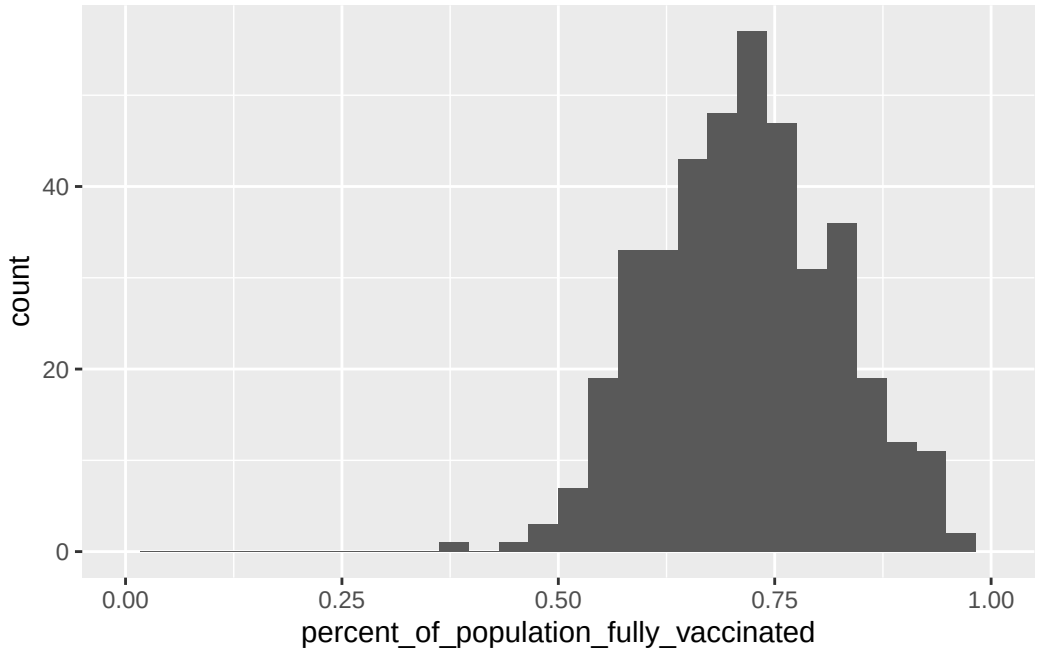
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3804  0.6457  0.7181  0.7213  0.7907  1.0000
```

**Q18.** Using ggplot generate a histogram of this data.

```
ggplot(vax.36)+
  aes(percent_of_population_fully_vaccinated)+
  geom_histogram()+
xlim(0,1)
```

```
`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?

The 92109 and 92040 ZIP code areas are above the average value.

```
x <- filter(vax.36, zip_code_tabulation_area %in% c("92109", "92040"))
x$percent_of_population_fully_vaccinated
```

```
[1] 0.694572 0.550296
```

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

```r
vax.36.all <- filter(vax, age5_plus_population > 36144)
ggplot(vax.36.all) +
  aes(as_of_date, percent_of_population_fully_vaccinated,
      percent_of_population_fully_vaccinated,
      group=zip_code_tabulation_area) +
  geom_line(alpha=0.2, color="blue") +
  labs(x="Date", y="Percent_Vaccinated",
       title="Vaccination_rate_across_California",
       subtitle="Only_areas_with_a_population_above_36k_are_shown") +
  geom_hline(yintercept = ave, linetype=2)
```

```
Warning: Removed 183 rows containing missing values (`geom_line()`).
```