Class 17: Mini-Project COVID-19 Vaccination Rates

Audrey Nguyen

Data Import

```
# import vaccination data
vax <- read.csv("29cd0b19-c7e6-4eb1-8be8-2b6e269f446e.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[nrow(vax)]
```

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

The earliest date in this dataset is 1/5/2021.

Q4. What is the latest date in this dataset?

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

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

The latest date in this dataset is 2/28/2023.

```
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
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	18993.87	0	1346.95	13685.13	31756.18	88556.7	
age5_plus_population	0	1.00	20875.24	21105.97	0	1460.50	15364.00	34877.00	101902.0	
tot_population	9718	0.95	23372.72	22628.51	12	2126.00	18714.00	38168.00	111165.0	
persons_fully_vaccinated	16525	0.92	13962.33	15054.09	11	930.00	8566.00	23302.00	87566.0	
persons_partially_vaccinated	16525	0.92	1701.64	2030.18	11	165.00	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	297.00	2748.00	9438.25	59553.0	
bivalent_dose_recip_count	158664	0.20	2924.93	3583.45	11	190.00	1418.00	4626.25	27458.0	
eligible_recipient_count	0	1.00	12801.84	14908.33	0	504.00	6338.00	21973.00	87234.0	

Q5. How many numeric columns are in this dataset?

There are 13 numeric columns.

Q6. How many `NA` values are 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. Why might this data be missing?

Not everyone might have reported their vaccination data.

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"
```

```
# this will give an error
# today() <- vax$as_of_date[1]
```

```
# specify that we're using the year-month-day format
vax$as_of_date <- ymd(vax$as_of_date)
```

How long does this 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() - ymd("2023-02-28")
```

```
Time difference of 7 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] 113
```

There are 113 unique dates in the dataset.

Working with ZIP codes

ZIP codes are also rather annoying 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 ZIP codes.

```
library(zipcodeR)
```

```
geocode_zip("92037")
```

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

```
zip_distance("92037", "95148")
```

```
  zipcode_a zipcode_b distance
1     92037     95148    405.6
```

```
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

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

```
[1] 12091
```

It is time to revisit the most awesome **dplyr** package.

```
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] 12091
```

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)
nrow(sd.10)
```

```
[1] 8588
```

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?

```
# find which zip code has the largest population
ind <- which.max(sd$age12_plus_population)
# display zip code by filtering
sd$zip_code_tabulation_area[ind]
```

```
[1] 92154
```

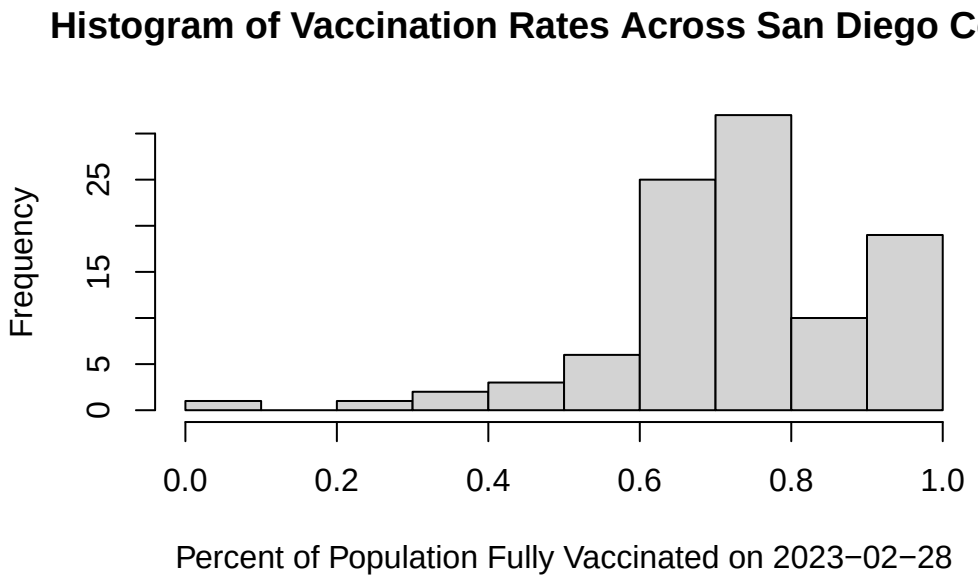
Q13. What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2023-02-28"?

```
sdtoday <- filter(sd, as_of_date == "2023-02-28")
mean(sdtoday$percent_of_population_fully_vaccinated, na.rm = TRUE)
```

```
[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 "2023-02-28"?

```
percent <- sdtoday$percent_of_population_fully_vaccinated
hist(percent, main = "Histogram of Vaccination Rates Across San Diego County", xlab = "Per
```



Focus on UCSD/La Jolla

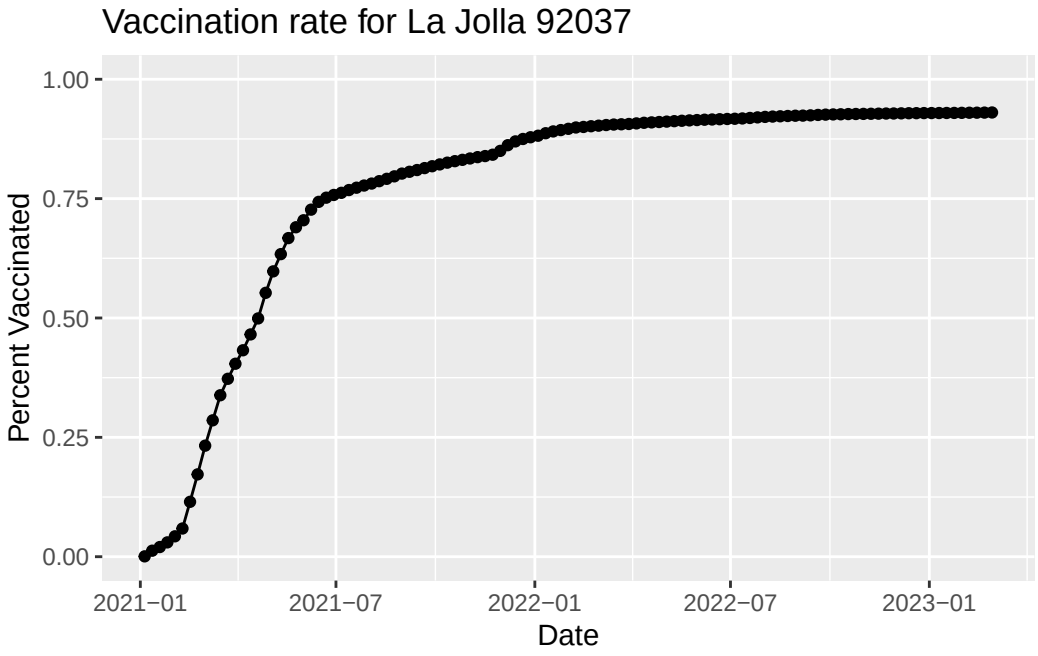
```
ucsd <- filter(sd, zip_code_tabulation_area == "92037")
ucsd[1, ]$age5_plus_population
```

```
[1] 36144
```

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

```
library(ggplot2)
```

```
ggplot(ucsd) +
  aes(as_of_date, percent_of_population_fully_vaccinated) + geom_point() + geom_line(group
```



Comparing to similar 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 == "2023-02-28")
head(vax.36)
```

```
  as_of_date zip_code_tabulation_area local_health_jurisdiction      county
1 2023-02-28                   93257                     Tulare      Tulare
2 2023-02-28                   93535                Los Angeles Los Angeles
3 2023-02-28                   91367                Los Angeles Los Angeles
4 2023-02-28                   90025                Los Angeles Los Angeles
5 2023-02-28                   90024                Los Angeles Los Angeles
6 2023-02-28                   90031                Los Angeles Los Angeles
  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                              4 Healthy Places Index Score
5                              3 Healthy Places Index Score
6                              1 Healthy Places Index Score
  age12_plus_population age5_plus_population tot_population
```

```
1               61519.8                70784          76519
2               59042.7                68471          74264
3               40437.4                43398          45970
4               42803.2                44982          46883
5               48841.8                50198          51627
6               34503.3                37735          39916
  persons_fully_vaccinated persons_partially_vaccinated
1                    45104                         5629
2                    45338                         4907
3                    33648                         2948
4                    36156                         4530
5                    28005                         5788
6                    29270                         3186
  percent_of_population_fully_vaccinated
1                               0.589448
2                               0.610498
3                               0.731956
4                               0.771196
5                               0.542449
6                               0.733290
  percent_of_population_partially_vaccinated
1                                   0.073563
2                                   0.066075
3                                   0.064129
4                                   0.096624
5                                   0.112112
6                                   0.079818
  percent_of_population_with_1_plus_dose booster_recip_count
1                               0.663011               22106
2                               0.676573               21799
3                               0.796085               22052
4                               0.867820               25207
5                               0.654561               19239
6                               0.813108               17344
  bivalent_dose_recip_count eligible_recipient_count redacted
1                      4981                    45046       No
2                      6754                    45247       No
3                      9234                    33544       No
4                     12099                    35980       No
5                      8578                    27934       No
6                      6076                    29213       No
```

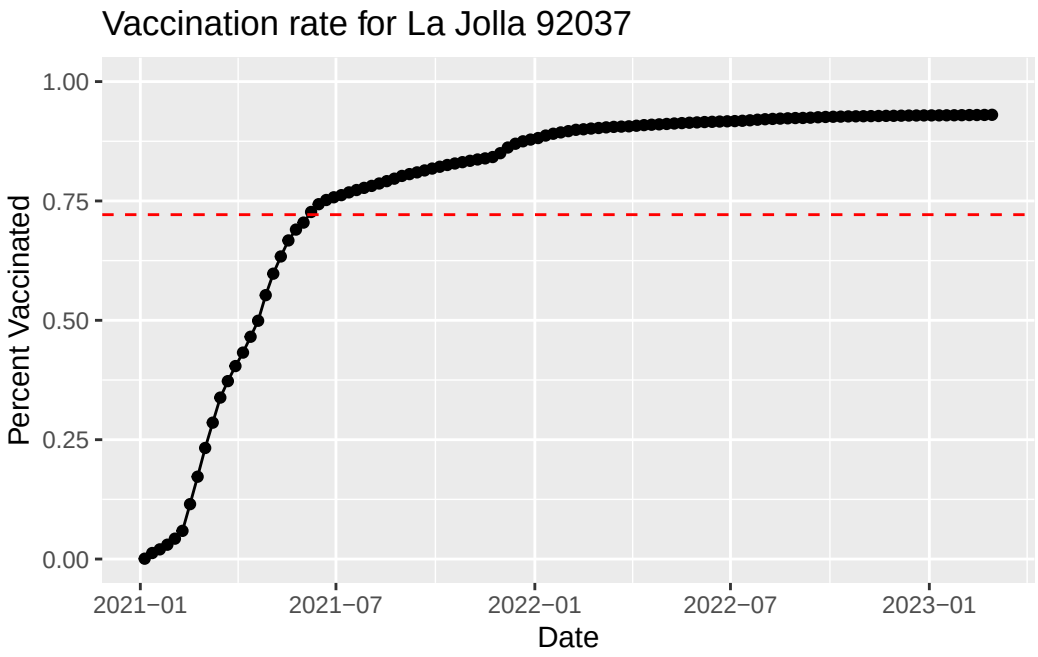
Q16. Calculate the mean "Percent of Population Fully Vaccinated" for ZIP code

areas with a population as large as 92037 (La Jolla) as_of_date "2023-02-28". Add this as as straight horizontal line to your plot from above with the `geom_hline()` function?

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

```
[1] 0.7213331
```

```
plot <- ggplot(ucsd) +
  aes(as_of_date, percent_of_population_fully_vaccinated) + geom_point() + geom_line(group
plot + geom_hline(aes(yintercept = avg), colour = "red", linetype = "dashed")
```



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 "2023-02-28"?

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

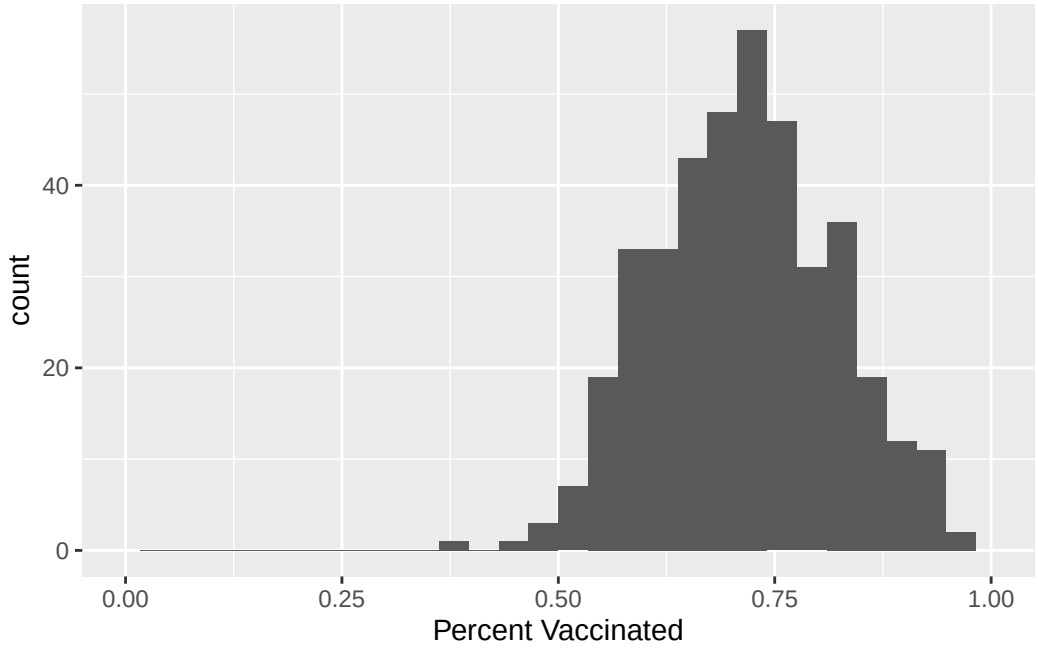
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3804  0.6457  0.7181  0.7213  0.7907  1.0000
```

Q18. Using ggplot, generate a histogram of this data:

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

```
`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?

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

```
[1] 0.694572 0.550296
```

The 92109 and 92040 values are below the average value for 92037.

Q20. Finally make a time course plot of vaccination progress for all areas in the full dataset 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,
      group=zip_code_tabulation_area) +
  geom_line(alpha=0.2, color="blue") +
  ylim(c(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 = avg, linetype= "dashed")
```

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