

Mini-Project: COVID-19 Vaccination Rates

Divya Shetty (A15390408)
3/3/2022

Getting Started

Import and examine the 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 92549 Riverside Riverside
## 2 2021-01-05 92130 San Diego San Diego
## 3 2021-01-05 92397 San Bernardino San Bernardino
## 4 2021-01-05 94563 Contra Costa Contra Costa
## 5 2021-01-05 94519 Contra Costa Contra Costa
## 6 2021-01-05 91042 Los Angeles Los Angeles
## vaccine_equity_metric_quartile vem_source
## 1 3 Healthy Places Index Score
## 2 4 Healthy Places Index Score
## 3 3 Healthy Places Index Score
## 4 4 Healthy Places Index Score
## 5 3 Healthy Places Index Score
## 6 2 Healthy Places Index Score
## age12_plus_population age5_plus_population persons_fully_vaccinated
## 1 2348.4 2461 NA
## 2 46300.3 53102 61
## 3 3695.6 4225 NA
## 4 17216.1 18896 NA
## 5 16861.2 18678 NA
## 6 23962.2 25741 NA
## persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1 NA NA
## 2 27 0.001149
## 3 NA NA
## 4 NA NA
## 5 NA NA
## 6 NA NA
## percent_of_population_partially_vaccinated
## 1 NA
## 2 0.000508
## 3 NA
## 4 NA
## 5 NA
## 6 NA
## percent_of_population_with_1_plus_dose booster_recip_count
## 1 NA NA
## 2 0.001657 NA
## 3 NA NA
## 4 NA NA
## 5 NA NA
## 6 NA NA
##
## 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?

The column is called "persons_fully_vaccinated".

Q2. What column details the Zip code tabulation area?

The column is "zip_code_tabulation_area".

Q3. What is the earliest date in this dataset?

```
vax$as_of_date[1]

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

The earliest date is "2021-01-05".

Q4. What is the latest date in this dataset?

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

## [1] "2022-03-01"
```

The latest date is "2022-03-01".

Use the skim() function to get an overview of the data.

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

Data summary

Name	vax
Number of rows	107604
Number of columns	15

Column type frequency:

character	5
numeric	10
<hr/>	
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	61	0
local_health_jurisdiction	0	1	0	15	305	62	0
county	0	1	0	15	305	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	5307	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	
age12_plus_population	0	1.00	18895.04	18993.91	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	18338	0.83	12155.61	13063.88	11	1066.25	7374.50	20005.00	77744.0	
persons_partially_vaccinated	18338	0.83	831.74	1348.68	11	76.00	372.00	1076.00	34219.0	
percent_of_population_fully_vaccinated	18338	0.83	0.51	0.26	0	0.33	0.54	0.70	1.0	
percent_of_population_partially_vaccinated	18338	0.83	0.05	0.09	0	0.01	0.03	0.05	1.0	
percent_of_population_with_1_plus_dose	18338	0.83	0.54	0.28	0	0.36	0.58	0.75	1.0	
booster_recip_count	64317	0.40	4100.55	5900.21	11	176.00	1136.00	6154.50	50602.0	

Q5. How many numeric columns are in this dataset?

There are 10 numeric columns.

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] 18338
```

There are 18338 NA values for "persons_fully_vaccinated".

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

```
sum(is.na(vax$persons_fully_vaccinated)) / length(vax$persons_fully_vaccinated)

## [1] 0.1704212
```

About 17.04% of the values in "persons_fully_vaccinated" are missing.

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

Missing values may be due to no records being collected from a given county.

Working with Dates

The lubridate package let's us use date data in a useful manner. Convert dates into lubridate formate and perform math operations with it!

```
library(lubridate)
today()

## [1] "2022-03-07"

#specify the year-month-day format when converting
vax$as_of_date <- ymd(vax$as_of_date)

#how many days have passed since the first vaccination?
today() - vax$as_of_date[1]

## Time difference of 426 days

#how many days does the dataset span?
vax$as_of_date[nrow(vax)] - vax$as_of_date[1]

## Time difference of 420 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
```

5 days have passed since the last update of the dataset.

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

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

```
## [1] 61
```

There are 61 different dates in the dataset.

Working with ZIP Codes

In R, we can use the `zipcodeR` package to make working with ZIP codes - a postal code used by the United States Postal Service (USPS) - easier.

```
library(zipcodeR)

#calculate the distance between the centroids of two zip codes in miles
zip_distance('92037','92109')

##   zipcode_a zipcode_b distance
## 1     92037     92109      2.33

#pull census data for zip code areas
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

Use the `dplyr` package to restrict the data to only the San Diego area.

```
library(dplyr)

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

```
## [1] 6527
```

The `dplyr` package can also be useful when trying to subset with multiple criteria. For example:

```
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
```

There are 107 unique zip codes for the San Diego County.

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

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

```
## [1] 92154
```

The '92154' zip code area has the largest 12+ population.

Select all San Diego "county" entries on "as_of_date" "2022-02-22" and use this for the following questions.

```
sd.feb <- filter(vax, county == "San Diego" & as_of_date == "2022-02-22")
```

Q13. What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2022-02-22"?

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

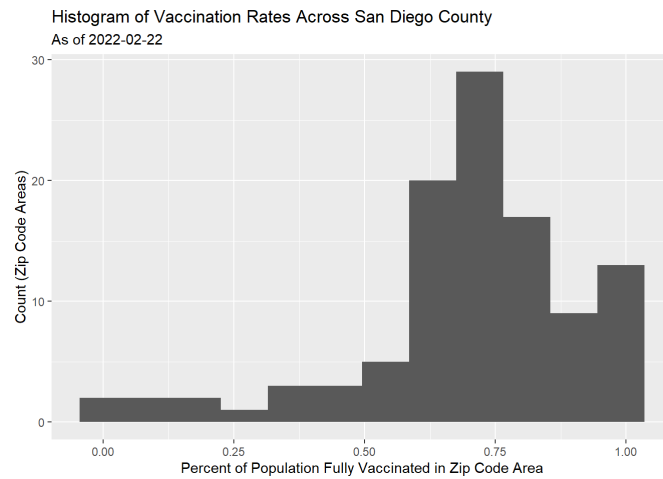
```
## [1] 0.7041551
```

The average percent of population fully vaccinated in the San Diego County as of 2022-02-22 is 70.42%.

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 "2022-02-22"?

```
library(ggplot2)

ggplot(sd.feb) +
  aes(x = percent_of_population_fully_vaccinated) +
  geom_histogram(bins = 12) +
  labs(title = "Histogram of Vaccination Rates Across San Diego County",
        subtitle = "As of 2022-02-22",
        x = "Percent of Population Fully Vaccinated in Zip Code Area",
        y = "Count (Zip Code Areas)")
```



Focus on UCSD/La Jolla

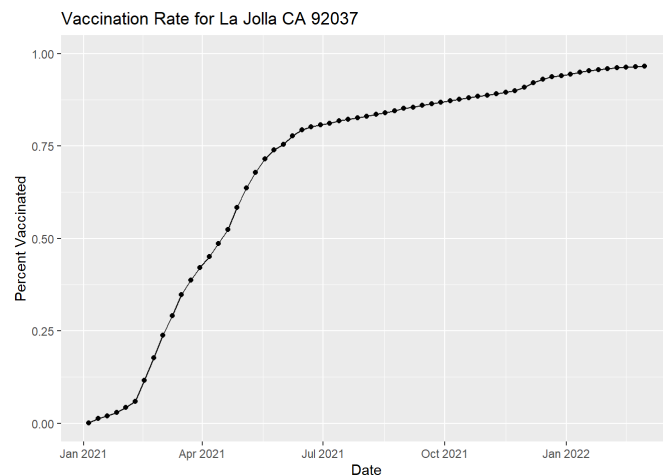
UC San Diego resides in the 92037 ZIP code area and is listed with an age 5+ population size of 36,144. Use this information to subset the data and check that it's done correctly.

```
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(x = as_of_date, y = percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group = 1) +
  ylim(c(0, 1)) +
  labs(title = "Vaccination Rate for La Jolla CA 92037",
       x = "Date", y = "Percent Vaccinated")
```



Comparing to Similar Size Areas

Look across every zip code area with a population at least as large as that of 92037 on as_of_date "2022-02-22".

```
vax.36 <- filter(vax, age5_plus_population > 36144 & as_of_date == "2022-02-22")
head(vax.36)
```

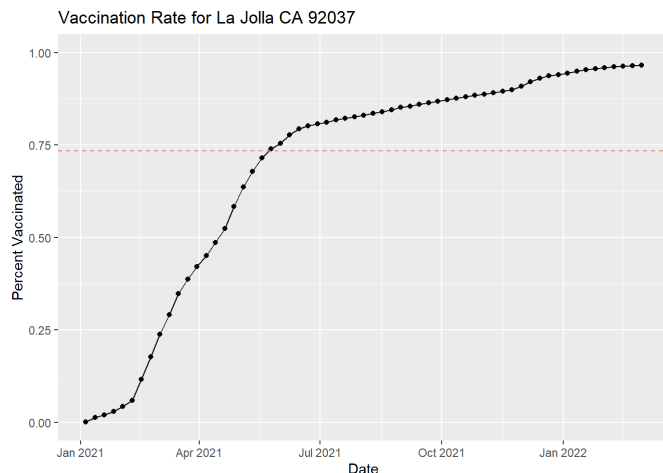
```
## as_of_date zip_code_tabulation_area local_health_jurisdiction county
## 1 2022-02-22 92840 Orange Orange
## 2 2022-02-22 92064 San Diego San Diego
## 3 2022-02-22 92508 Riverside Riverside
## 4 2022-02-22 95403 Sonoma Sonoma
## 5 2022-02-22 90001 Los Angeles Los Angeles
## 6 2022-02-22 92802 Orange Orange
## vaccine_equity_metric_quartile vem_source
## 1 2 Healthy Places Index Score
## 2 4 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 47302.5 51902 40725
## 2 42177.1 46855 34266
## 3 32415.3 36303 21925
## 4 38545.9 42294 33158
## 5 47175.7 54805 43075
## 6 35113.6 39393 29268
## persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1 4324 0.784652
## 2 6861 0.731320
## 3 1714 0.603945
## 4 2833 0.783988
## 5 13917 0.785968
## 6 6138 0.742975
## percent_of_population_partially_vaccinated
## 1 0.083311
## 2 0.146430
## 3 0.047214
## 4 0.066983
## 5 0.253937
## 6 0.155814
## percent_of_population_with_1_plus_dose booster_recip_count redacted
## 1 0.867963 20654 No
## 2 0.877750 15499 No
## 3 0.651159 10753 No
## 4 0.850971 18659 No
## 5 1.000000 13408 No
## 6 0.898789 12816 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 "2022-02-22". Add this as a straight horizontal line to your plot from above with the geom_hline() function?

```
avg.per <- mean(vax.36$percent_of_population_fully_vaccinated, na.rm = TRUE)
avg.per
```

```
## [1] 0.733385
```

```
ggplot(ucsd) +
  aes(x = as_of_date, y = percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group = 1) +
  ylim(c(0, 1)) +
  labs(title = "Vaccination Rate for La Jolla CA 92037",
       x = "Date", y = "Percent Vaccinated") +
  geom_hline(aes(yintercept = avg.per, color = "red"),
            linetype = "dashed", show.legend = FALSE)
```



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 "2022-02-22"?

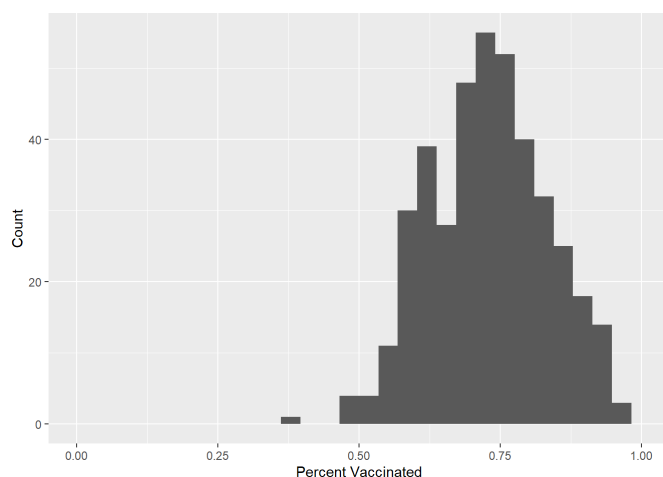
```
summary(vax.36$percent_of_population_fully_vaccinated)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.3881  0.6539  0.7333  0.7334  0.8027  1.0000
```

Q18. Using ggplot generate a histogram of this data.

```
ggplot(vax.36) +
  aes(x = percent_of_population_fully_vaccinated) +
  geom_histogram(bins = 30) +
  labs(x = "Percent Vaccinated", y = "Count") +
  xlim(c(0, 1))
```

```
## 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?

```
#92109
vax %>% filter(as_of_date == "2022-02-22") %>%
  filter(zip_code_tabulation_area=="92109") %>%
  select(percent_of_population_fully_vaccinated)
```

```
## percent_of_population_fully_vaccinated
## 1 0.723044
```

```
#92040
vax %>% filter(as_of_date == "2022-02-22") %>%
  filter(zip_code_tabulation_area == "92040") %>%
  select(percent_of_population_fully_vaccinated)
```

```
## percent_of_population_fully_vaccinated
## 1 0.551304
```

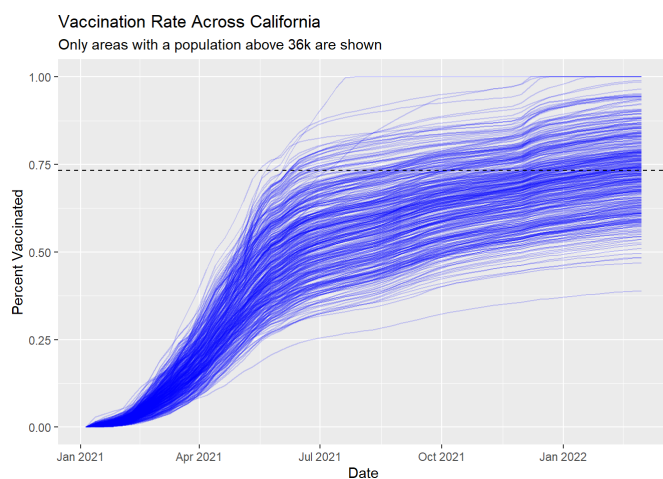
Both zip code areas are below the average value calculated.

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

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

ggplot(vax.36.all) +
  aes(x = as_of_date,
      y = 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.per, linetype = "dashed")
```

```
## Warning: Removed 311 row(s) containing missing values (geom_path).
```



Q21. How do you feel about traveling for Spring Break and meeting for in-person class afterwards?

I would not feel comfortable traveling for break considering that nearly half of the largely populated areas don't have a great majority of their population vaccinated. This would make having in-person classes after break riskier since we can't guarantee that each location everyone traveled to was safe.