Optional Vaccination Rate Project

Julia Ainsworth

Table of contents

Beginning:

Importing vaccination data for San Diego:

```
vax <- read.csv(file = "covid19vaccinesbyzipcode_test.csv")
# head(vax) note: was cluttering up pdf
```

Q1. What column details the total number of people fully vaccinated?

A: persons_fully_vaccinated

Q2. What column details the Zip code tabulation area?

A: 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"
```

```
tail(vax$as_of_date)
```

```
[1] "2022-11-22" "2022-11-22" "2022-11-22" "2022-11-22" "2022-11-22"
[6] "2022-11-22"
```

Q3. What is the earliest date in this dataset?

A: 2021-01-05

Q4. What is the latest date in this dataset?

A: 2022-11-22

```
# loaded skimr library in the console
skimr::skim(vax)
```

Table 1: Data summary

Name	vax
Number of rows	174636
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	99	0
local_health_jurisdiction	0	1	0	15	495	62	0
county	0	1	0	15	495	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_quintile	8613	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	
age12_plus_population	0	1.00	18895.04	18993.88	0	1346.95	13685.10	31756.12	88556.7	
age5_plus_population	0	1.00	20875.24	21105.98	0	1460.50	15364.00	34877.00	101902.0	
tot_population	8514	0.95	23372.77	22628.51	12	2126.00	18714.00	38168.00	111165.0	
persons_fully_vaccinated	14921	0.91	13466.34	14722.46	11	883.00	8024.00	22529.00	87186.0	
persons_partially_vaccinated	14921	0.91	1707.50	1998.80	11	167.00	1194.00	2547.00	39204.0	
percent_of_population_fully_vaccinated	18665	0.89	0.55	0.25	0	0.39	0.59	0.73	1.0	
percent_of_population_partially_vaccinated	18665	0.89	0.08	0.09	0	0.05	0.06	0.08	1.0	
percent_of_population_with_1_plus_dose	19562	0.89	0.61	0.25	0	0.46	0.65	0.79	1.0	
booster_recip_count	70421	0.60	5655.17	6867.49	11	280.00	2575.00	9421.00	58304.0	
bivalent_dose_recip_count	156958	0.10	1646.02	2161.84	11	109.00	719.00	2443.00	18109.0	
eligible_recipient_count	0	1.00	12309.19	14555.83	0	466.00	5810.00	21140.00	86696.0	

Q5. How many numeric columns are in this dataset?

A: 13

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

14921 (code below)

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

.00069 %

```r
a <- sum( is.na(vax$persons_fully_vaccinated) ) #Question 6
a
```

```
[1] 14921
```

```r
b <- sum(vax$persons_fully_vaccinated, na.rm = T)

(a/b) * 100 # Question 7
```

```
[1] 0.0006937493
```

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

They may have been vaccinated originally in a different county, so not all their records are available in San Diego

Working with dates

```r
library(lubridate)
```

```
Loading required package: timechange


Attaching package: 'lubridate'


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

    date, intersect, setdiff, union
```

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

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

```
Time difference of 9 days
```

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

```
nrow(table(vax$as_of_date))
```

```
[1] 99
```

Working with zip codes

```
library(zipcodeR)
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
```

Using dplyr to look at San Diego vaccinations:

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

nrow(sd)
```

```
[1] 10593
```

Filtering for areas with population of over 10000:

```
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] 53
```

```
sd$zip_code_tabulation_area[53]
```

```
[1] 92154
```

Q13. What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2022-11-15"?

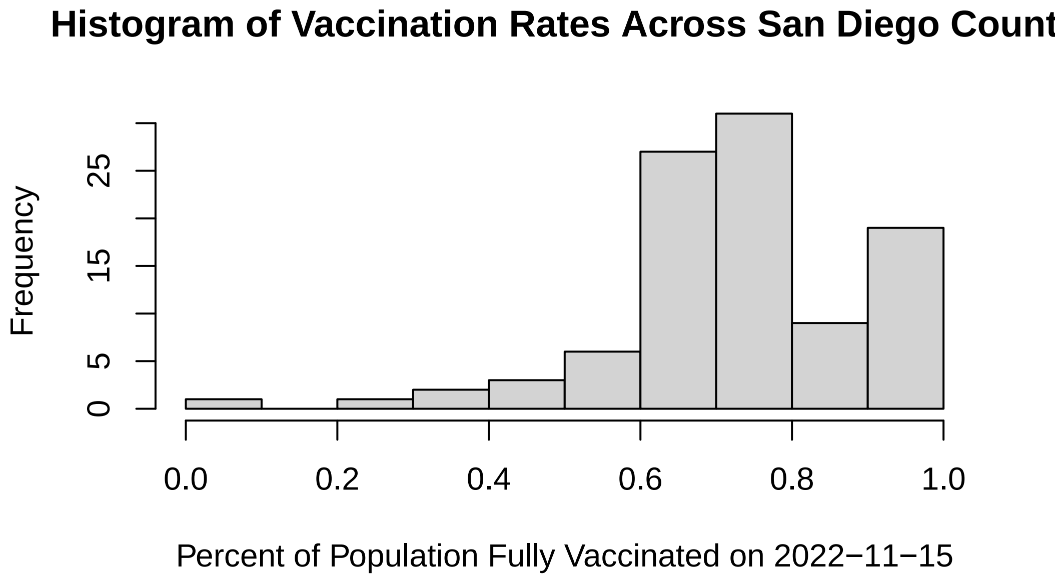
Filtering with dplyr:

```
sd.subset <- filter(vax, county == "San Diego" &
                as_of_date == "2022-11-15" )

mean.default(sd.subset$percent_of_population_fully_vaccinated, na.rm = T)*100
```

```
[1] 73.69099
```

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.data <- sd.subset$percent_of_population_fully_vaccinated
hist(hist.data, main = "Histogram of Vaccination Rates Across San Diego County",
     xlab = "Percent of Population Fully Vaccinated on 2022-11-15")
```



Narrowing in on only La Jolla:

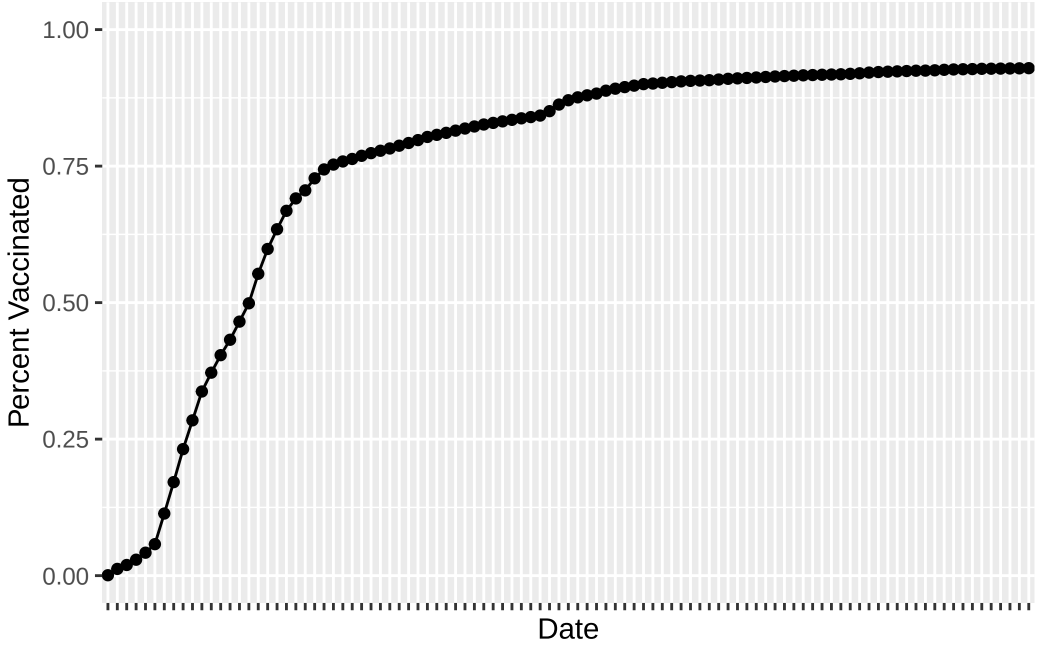
```
ucsd <- filter(sd, zip_code_tabulation_area=="92037")
ucsd[1,]$age5_plus_population
```

```
[1] 36144
```

```
# head(ucsd)
```

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

```
library(ggplot2)
plot.a <-  ggplot(ucsd) +
  aes(x = as_of_date,
      percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) +
  ylim(c(0,1)) +
  labs(x = "Date", y="Percent Vaccinated") + theme(axis.text.x = element_blank())
plot.a
```



# Subset to all CA areas with a population as large as 92037

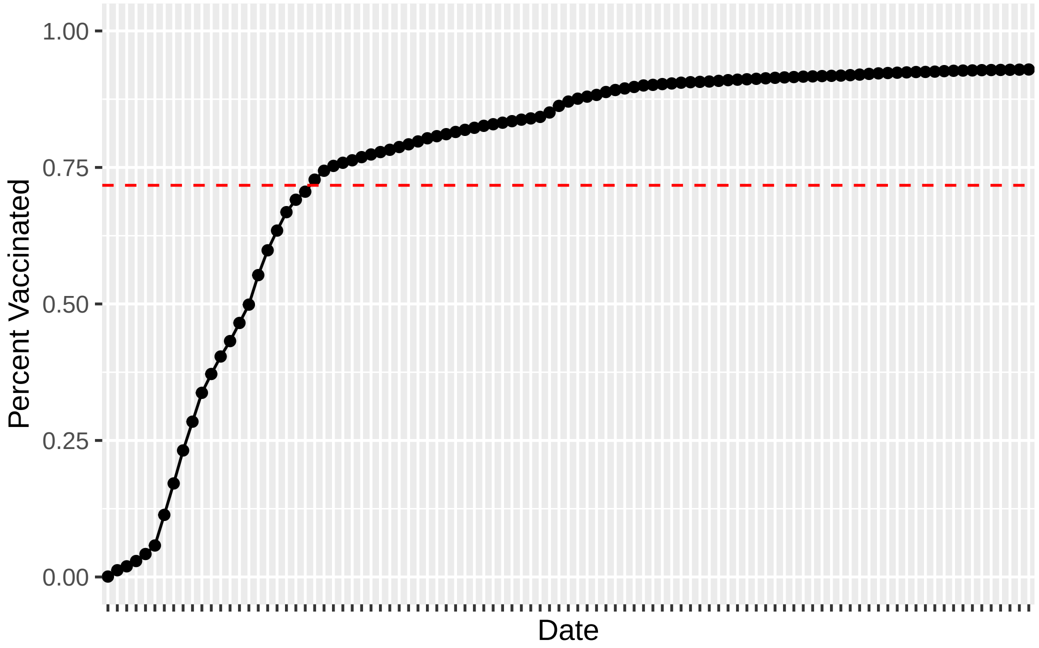
```
vax.36 <- filter(vax, age5_plus_population > 36144 &
                    as_of_date == "2022-11-15")

# 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?

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

```
[1] 0.7172851
```

```
plot.a +
  geom_hline(yintercept = mean.36, 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"?

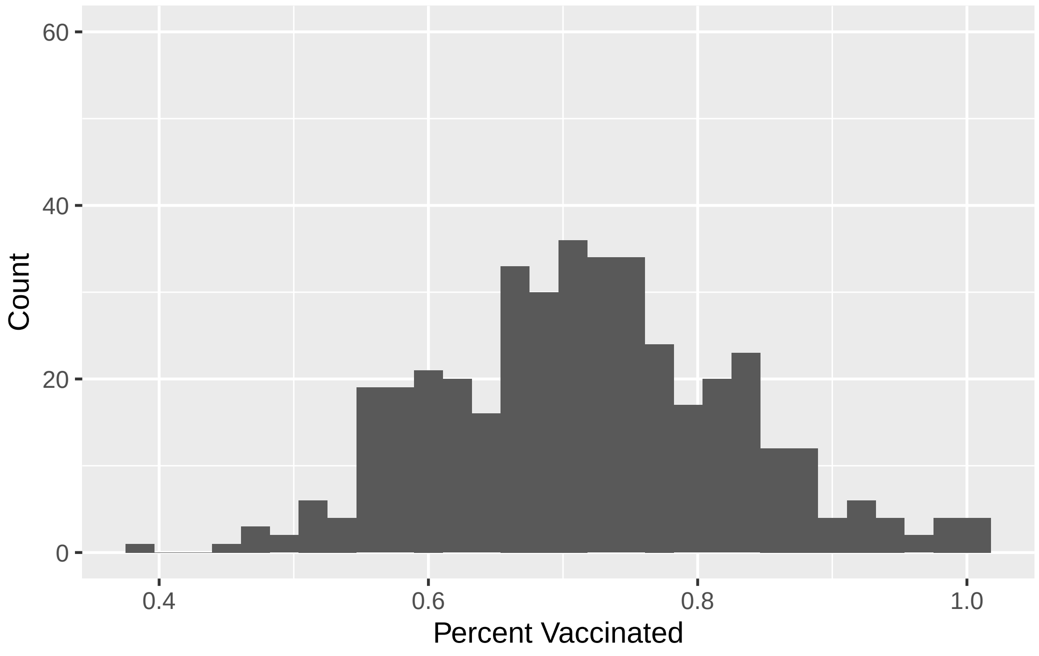
```
summary(vax.36$percent_of_population_fully_vaccinated)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.3785  0.6396  0.7155  0.7173  0.7880  1.0000 
```

Q18. Using ggplot generate a histogram of this data.

```
ggplot(vax.36) +
  aes(percent_of_population_fully_vaccinated) +
  geom_histogram() + xlab("Percent Vaccinated") + ylab("Count") + ylim(0,60)
```

```
`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 == "2022-11-15") %>%
  filter(zip_code_tabulation_area=="92040") %>%
  select(percent_of_population_fully_vaccinated)
```

```
  percent_of_population_fully_vaccinated
1                               0.546646
```

```
vax %>% filter(as_of_date == "2022-11-15") %>%
  filter(zip_code_tabulation_area=="92109") %>%
  select(percent_of_population_fully_vaccinated)
```

```
  percent_of_population_fully_vaccinated
1                               0.693299
```

92040: below 92109: below

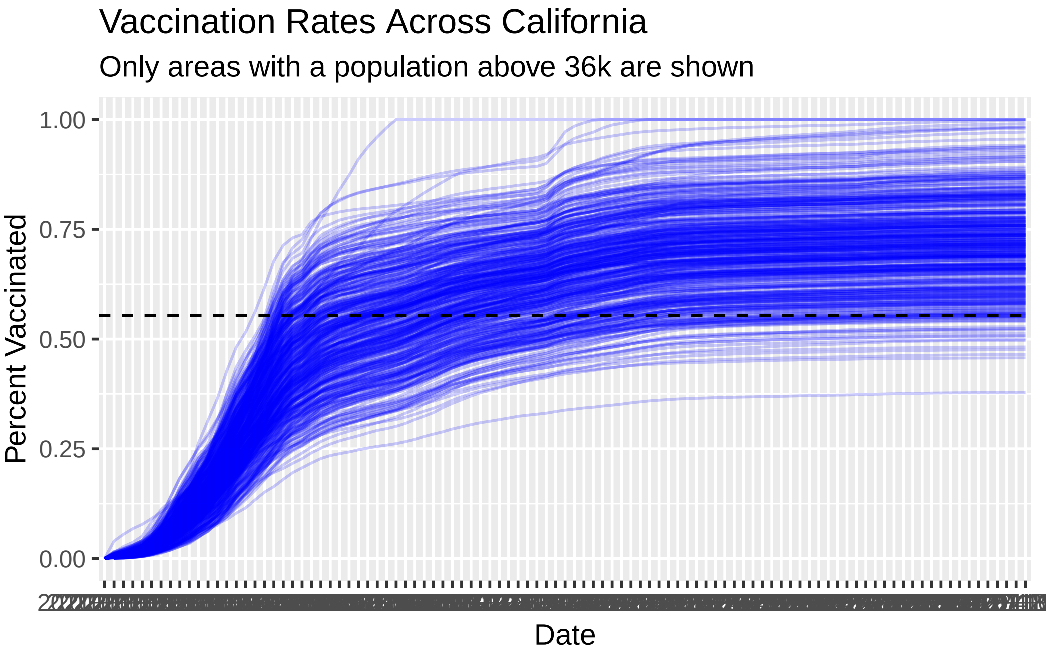
Q20. Finally make a time course plot of vaccination progress for all areas in the full dataset with a age5_plus_population $>$ 36144.

```
vax.36.all <- filter(vax, age5_plus_population > 36144)
# head(vax.36.all)
mean.36.all <- mean.default(vax.36.all$percent_of_population_fully_vaccinated, na.rm = T)
mean.36.all*100
```

```
[1] 55.34134
```

```
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 Rates Across California",
       subtitle="Only areas with a population above 36k are shown") +
  geom_hline(yintercept = mean.36.all, linetype = 2)
```

```
Warning: Removed 184 row(s) containing missing values (geom_path).
```



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

I didn't travel for Thanksgiving, but I will test before coming to class. I really prefer in-person class.