Vaccination Mini Project

Ralph Goguanco

3/11/2022

The goal of this hands-on mini-project is to examine and compare the Covid-19 vaccination rates around San Diego.

```
# Import vaccination data
vax <- read.csv("covid19vaccinesbyzipcode_test.csv")
head(vax)
```

	as_of_date <chr>	zip_code_tabulation_area <int>	local_health_jurisdiction <chr>	county <chr>
1	2021-01-05	95959	Nevada	Nevada
2	2021-01-05	95694	Yolo	Yolo
3	2021-01-05	95714	Placer	Placer
4	2021-01-05	95843	Sacramento	Sacramento
5	2021-01-05	95935	Yuba	Yuba
6	2021-01-05	95970	Colusa	Colusa

6 rows | 1-5 of 16 columns

> 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?

```
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"
```

```
#2021-01-05
```

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

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

```
## [1] "2022-03-08" "2022-03-08" "2022-03-08" "2022-03-08" "2022-03-08"
## [6] "2022-03-08"
```

```
#2022-03-08
```

As we have done previously, let's call the skim() function from the skimr package to get a quick overview of this dataset:

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

Data summary

Name	vax

Number of rows	109368
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	62	0
local_health_jurisdiction	0	1	0	15	310	62	0
county	0	1	0	15	310	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	5394	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	21106.01	0	1460.50	15364.00	34877.00	101902.0	▇▂▂▁
persons_fully_vaccinated	18494	0.83	12246.65	13155.33	11	1074.00	7453.00	20138.00	79763.0	▇▂▂▁
persons_partially_vaccinated	18494	0.83	848.69	1401.05	11	77.00	377.00	1091.00	36844.0	▇▁▁▁
percent_of_population_fully_vaccinated	18494	0.83	0.51	0.26	0	0.34	0.55	0.71	1.0	▃▅▆▇
percent_of_population_partially_vaccinated	18494	0.83	0.05	0.10	0	0.02	0.03	0.05	1.0	▇▁▁▁
percent_of_population_with_1_plus_dose	18494	0.83	0.55	0.28	0	0.36	0.59	0.76	1.0	▃▅▆▇
booster_recip_count	64441	0.41	4265.63	6073.54	11	183.00	1207.00	6484.00	51001.0	▇▂▁▁

Q5. How many numeric columns are in this dataset?

9 >Q6. Note that there are “missing values” in the dataset. How many NA values there in the persons_fully_vaccinated column?

```
N <- sum(is.na(vax$persons_fully_vaccinated))
total <- sum(vax$percent_of_population_fully_vaccinated, na.rm = TRUE)
```

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

```
N/total
```

```
## [1] 0.3959551
```

#Working with dates

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
today()
```

```
## [1] "2022-03-11"
```

```
#Specify that we are using the year-month-day format
vax$as_of_date <- ymd(vax$as_of_date)
```

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

```
## Time difference of 430 days
```

Q9. How many days have passed since the last update of the dataset? 3 days

```
today() - vax$as_of_date[109358]
```

```
## Time difference of 3 days
```

Q10. How many unique dates are in the dataset (i.e. how many different dates are detailed)? 62

```
date_uniq <- unique(vax$as_of_date)
length(date_uniq)
```

```
## [1] 62
```

#Working with ZIP codes

```
library(zipcodeR)
# Pull data for all ZIP codes in the dataset
zipdata <- reverse_zipcode( vax$zip_code_tabulation_area )
```

```
# Subset to San Diego county only areas
sd <- vax[ 4 , ]
```

```
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] 6634
```

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

```
length(unique(sd))
```

```
## [1] 15
```

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

```
which.max(sd[, 7])
```

```
## [1] 103
```

```
# Zip Code in #103 is 92154
```

```
sd1 <- filter(sd, 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”?

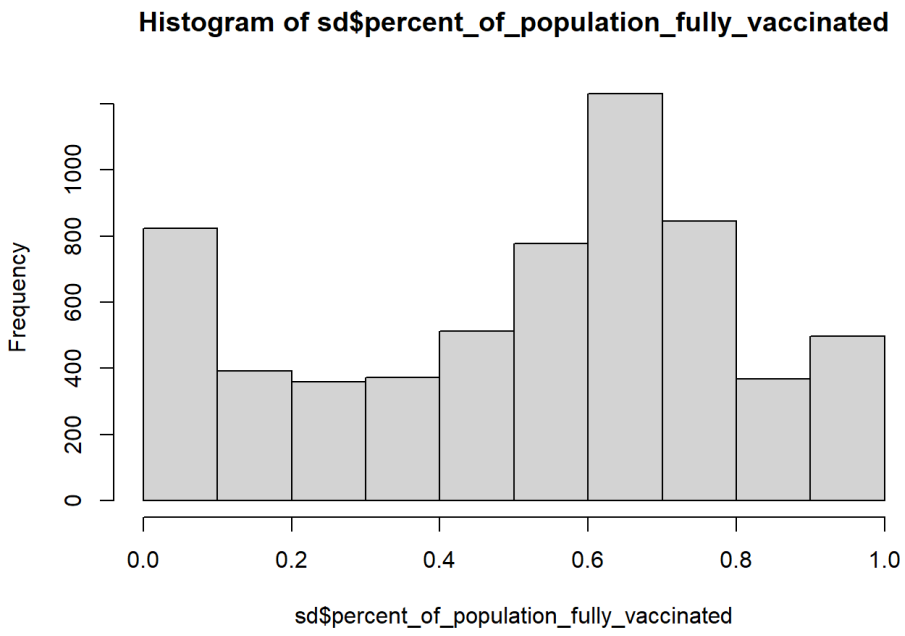
```
mean(sd1$percent_of_population_fully_vaccinated, na.rm = TRUE)
```

```
## [1] 0.7145899
```

71.46%

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”?

```
hist(sd$percent_of_population_fully_vaccinated)
```



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:

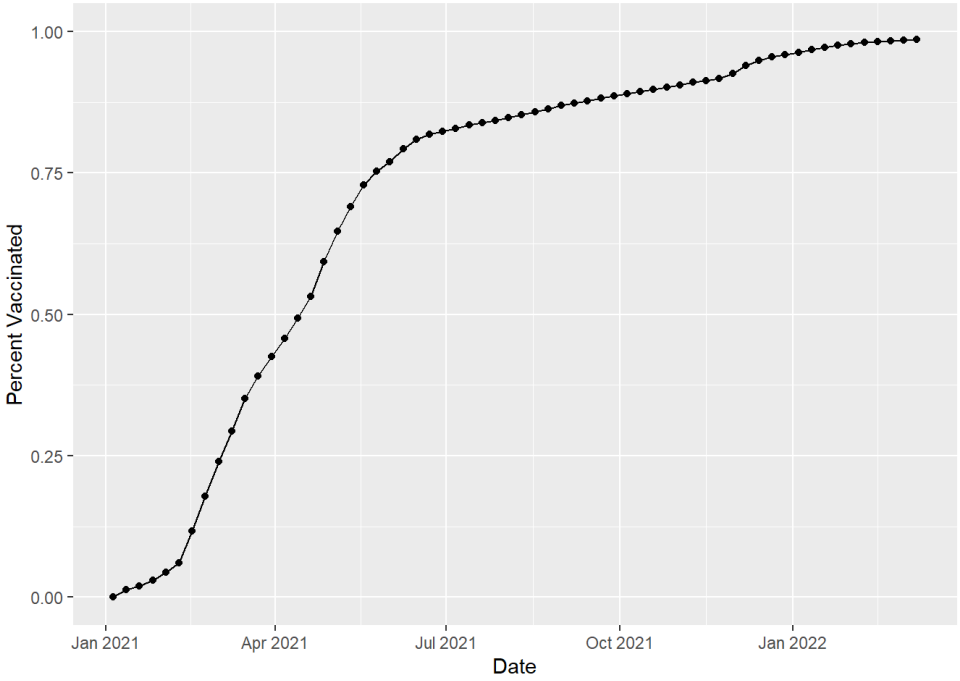
```
library(ggplot2)
ggplot(ucsd) +
  aes(x=ucsd$as_of_date,
      y = ucsd$percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) +
  ylim(c(0,1)) +
  labs(x="Date", y="Percent Vaccinated")
```

```
## Warning: Use of `ucsd$as_of_date` is discouraged. Use `as_of_date` instead.
```

```
## Warning: Use of `ucsd$percent_of_population_fully_vaccinated` is discouraged.
## Use `percent_of_population_fully_vaccinated` instead.
```

```
## Warning: Use of `ucsd$as_of_date` is discouraged. Use `as_of_date` instead.
```

```
## Warning: Use of `ucsd$percent_of_population_fully_vaccinated` is discouraged.
## Use `percent_of_population_fully_vaccinated` instead.
```



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 == "2022-02-22")

head(vax.36)
```

	as_of_date <date>	zip_code_tabulation_area <int>	local_health_jurisdiction <chr>	county <chr>
1	2022-02-22	94110	San Francisco	San Francisco

	as_of_date <date>	zip_code_tabulation_area <int>	local_health_jurisdiction <chr>	county <chr>
2	2022-02-22	93552	Los Angeles	Los Angeles
3	2022-02-22	93611	Fresno	Fresno
4	2022-02-22	92336	San Bernardino	San Bernardino
5	2022-02-22	95116	Santa Clara	Santa Clara
6	2022-02-22	92571	Riverside	Riverside

6 rows | 1-5 of 16 columns

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?

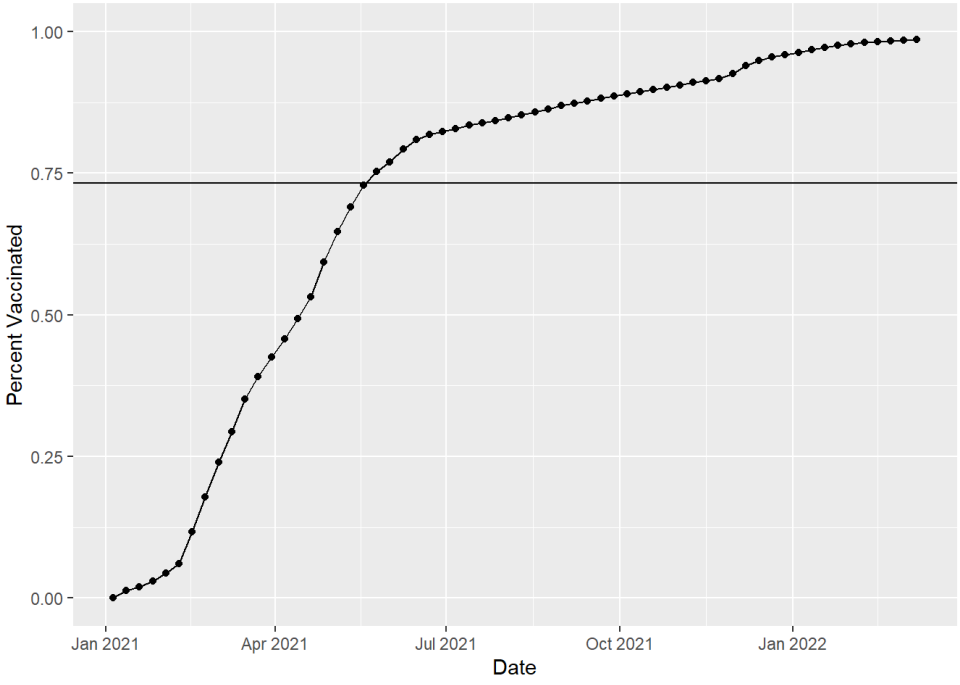
```
ggplot(ucsd) +
aes(x=ucsd$as_of_date,
      y = ucsd$percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) +
  ylim(c(0,1)) +
  labs(x="Date", y="Percent Vaccinated") +
  geom_hline(yintercept = 0.732736)
```

```
## Warning: Use of `ucsd$as_of_date` is discouraged. Use `as_of_date` instead.
```

```
## Warning: Use of `ucsd$percent_of_population_fully_vaccinated` is discouraged.
## Use `percent_of_population_fully_vaccinated` instead.
```

```
## Warning: Use of `ucsd$as_of_date` is discouraged. Use `as_of_date` instead.
```

```
## Warning: Use of `ucsd$percent_of_population_fully_vaccinated` is discouraged.
## Use `percent_of_population_fully_vaccinated` instead.
```



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”?

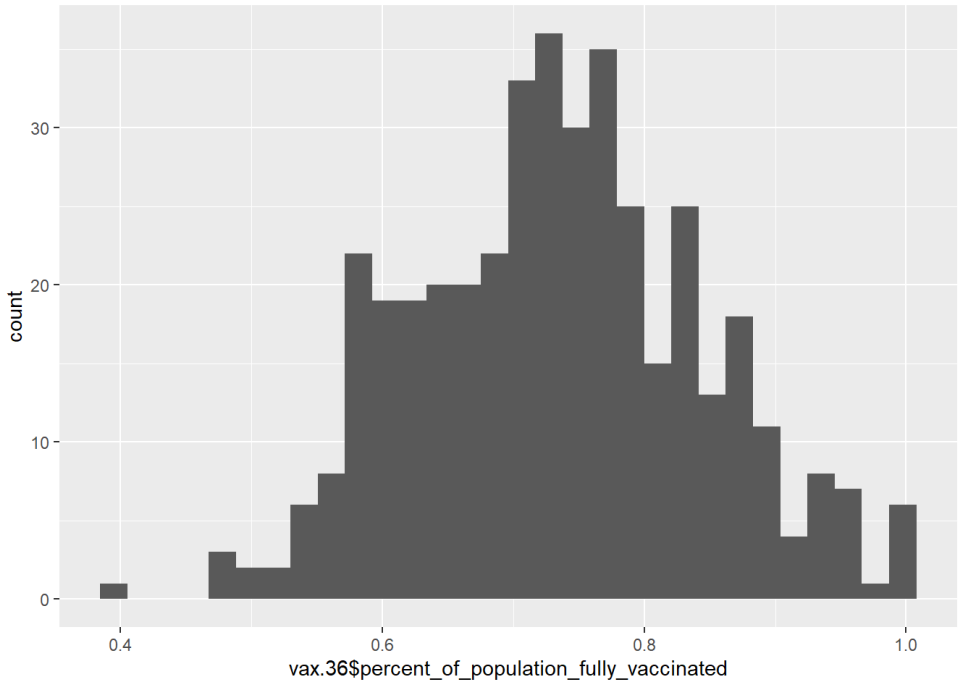
```
#fivenum(vax.36)
```

Q18. Using ggplot generate a histogram of this data.

```
ggplot(vax.36, aes(x=vax.36$percent_of_population_fully_vaccinated)) +
geom_histogram()
```

```
## Warning: Use of `vax.36$percent_of_population_fully_vaccinated` is discouraged.
## Use `percent_of_population_fully_vaccinated` instead.
```

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

percent_of_population_fully_vaccinated
<dbl>
0.559102

1 row

Below the average value

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)


#ggplot(vax.36.all) +
# aes(x = vax.36$as_of_date,
#      percent_of_population_fully_vaccinated,
#      group=zip_code_tabulation_area) +
#  geom_line(alpha=0.2, color=vax) +
#  ylim(vax.36) +
# labs(x="Date" , y="Percent",
#       title="Vaccination rate accross California",
#       subtitle="Only areas with a population above 36k are shown") +
#  geom_hline(yintercept = 0.7327, linetype=dash)
```

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

I feel better knowing that the percent vaccinated will only go up