Class#17

Zainab Ashir

2023-06-03

```
# Import vaccination data
data <- "https://marcos-diazg.github.io/BIMM143_SP23/class-material/class17/covid19vaccinesbyzipcode_test.csv"
vax <- read.csv(data)
head(vax)
```

	as_of_date <chr>	zip_code_tabulation_area <int>	local_health_jurisdiction <chr>	county <chr>
1	2021-01-05	94579	Alameda	Alameda
2	2021-01-05	93726	Fresno	Fresno
3	2021-01-05	94305	Santa Clara	Santa Clara
4	2021-01-05	93704	Fresno	Fresno
5	2021-01-05	94403	San Mateo	San Mateo
6	2021-01-05	93668	Fresno	Fresno

6 rows | 1-5 of 20 columns

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

##It would be the column labeled “persons_fully_vaccinated”

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

##It would be the column labeled “zip_code-tabulation_area”

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

##The earliest date is 2021-01-05

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

##The latest date is 2023-05-23

```
skimr::skim_without_charts(vax)
```

Data summary

Name	vax
Number of rows	220500
Number of columns	19
_____	
Column type frequency:	
character	5
numeric	14

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	125	0
local_health_jurisdiction	0	1	0	15	625	62	0
county	0	1	0	15	625	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
zip_code_tabulation_area	0	1.00	93665.11	1817.38	90001	92257.75	93658.50	95380.50	97635.0
vaccine_equity_metric_quartile	10875	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0
age12_plus_population	0	1.00	18895.04	18993.87	0	1346.95	13685.10	31756.12	88556.7
age5_plus_population	0	1.00	20875.24	21105.97	0	1460.50	15364.00	34877.00	101902.0
tot_population	10750	0.95	23372.77	22628.50	12	2126.00	18714.00	38168.00	111165.0
persons_fully_vaccinated	17711	0.92	14272.72	15264.17	11	954.00	8990.00	23782.00	87724.0
persons_partially_vaccinated	17711	0.92	1711.05	2071.56	11	164.00	1203.00	2550.00	42259.0
percent_of_population_fully_vaccinated	22579	0.90	0.58	0.25	0	0.44	0.62	0.75	1.0
percent_of_population_partially_vaccinated	22579	0.90	0.08	0.09	0	0.05	0.06	0.08	1.0
percent_of_population_with_1_plus_dose	23732	0.89	0.64	0.24	0	0.50	0.68	0.82	1.0
booster_recip_count	74388	0.66	6373.43	7751.70	11	328.00	3097.00	10274.00	60022.0
bivalent_dose_recip_count	159956	0.27	3407.91	4010.38	11	222.00	1832.00	5482.00	29484.0
eligible_recipient_count	0	1.00	13120.40	15126.17	0	534.00	6663.00	22517.25	87437.0
eligible_bivalent_recipient_count	0	1.00	13016.51	15199.08	0	266.00	6562.00	22513.00	87437.0

Q5. How many numeric columns are in this dataset?

##There are 14 numeric colums.

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

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

```
## [1] 17711
```

##Thus, there is 17711 NA values.

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

```
sum(is.na(vax$persons_fully_vaccinated)) / nrow(vax) * 100
```

```
## [1] 8.0322
```

##Thus, it seems that percent of persons_fully_vaccinated is 8.03%

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

```
library(lubridate)
```

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

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

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

```
## Time difference of 11 days
```

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

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

```
## [1] 125
```

##Thus, there are 125 unique dates in the dataset.

```
library(zipcodeR)
```

```
## The legacy packages maptools, rgdal, and rgeos, underpinning this package
## will retire shortly. Please refer to R-spatial evolution reports on
## https://r-spatial.org/r/2023/05/15/evolution4.html for details.
## This package is now running under evolution status 0
```

```
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
```

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

zipcode	lat	lng
<chr>	<dbl>	<dbl>
92037	32.8	-117.2

1 row

```
zip_distance('92037','92109')
```

zipcode_a	zipcode_b	distance
<chr>	<chr>	<dbl>
92037	92109	2.33

1 row

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

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

```
## [1] 13375
```

```
sd.zip <- filter(vax, county == "San Diego" &
                 vax$zip_code_tabulation_area)
SD_unique_zip<- length(unique(sd.zip))
SD_unique_zip
```

```
## [1] 19
```

##It seems that there are 19 unique zip codes listed for san diego.

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

```
which.max(sd.zip$tot_population)
```

```
## [1] 87
```

```
sd.zip
```

as_of_date	zip_code_tabulation_area	local_health_jurisdiction	county
<date>	<int>	<chr>	<chr>
2021-01-05	91977	San Diego	San Diego
2021-01-05	92110	San Diego	San Diego
2021-01-05	92101	San Diego	San Diego

as_of_date <date>	zip_code_tabulation_area <int>	local_health_jurisdiction <chr>	county <chr>
2021-01-05	92071	San Diego	San Diego
2021-01-05	92070	San Diego	San Diego
2021-01-05	92028	San Diego	San Diego
2021-01-05	92024	San Diego	San Diego
2021-01-05	92059	San Diego	San Diego
2021-01-05	92021	San Diego	San Diego
2021-01-05	92003	San Diego	San Diego

1-10 of 10,000 rows | 1-4 of 19 columns

Previous **1** 2 3 4 5 6 … 1000 Next

##Thus, this 87 number of row corrosponds to 92154 zipcode.

Q13. What is the overall average (with 2 decimal numbers) “Percent of Population Fully Vaccinated” value for all San Diego “County” as of “2023-05-23”?

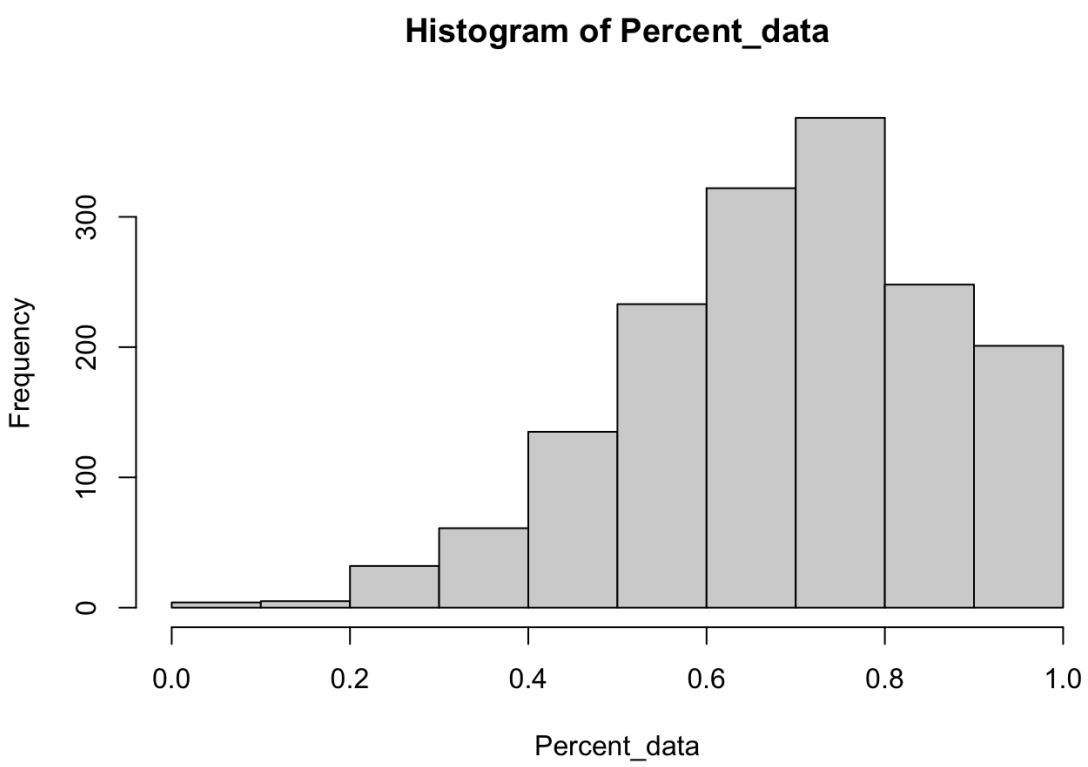
```
SD_data <- vax[vax$county == "San Diego" & vax$as_of_date == "2023-05-23", ]
average_percent_vaccinated <- mean(SD_data$percent_of_population_fully_vaccinated, na.rm = TRUE)
average_percent_vaccinated * 100
```

```
## [1] 74.19654
```

##Thus, it seems that the answer is 74.20%

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 “2023-05-23”?

```
date_data <- vax[vax$as_of_date == "2023-05-23", ]
Percent_data<-date_data$percent_of_population_fully_vaccinated
hist(Percent_data)
```

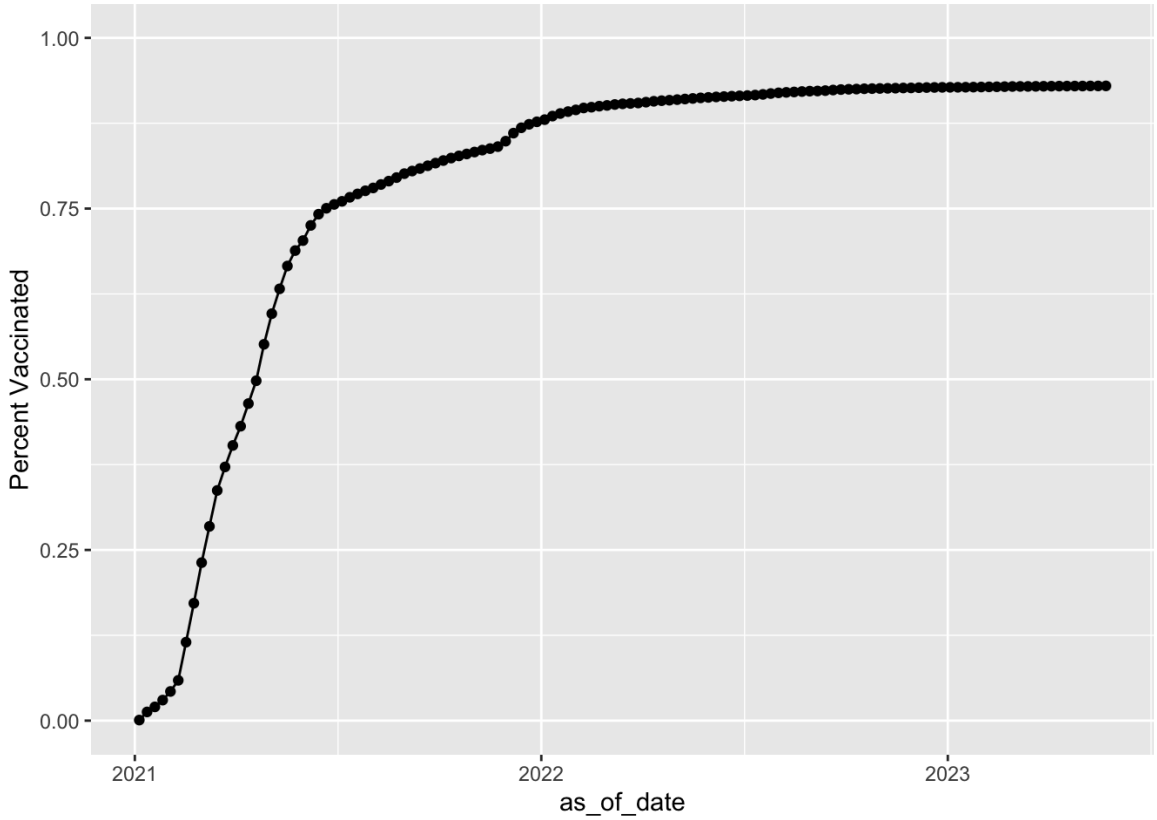


Q15. Using ggplot make a graph of the vaccination rate time course for the 92037 ZIP code area

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

```
## [1] 36144
```

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



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 “2023-05-23”. Add this as a straight horizontal line to your plot from above with the geom_hline() function?

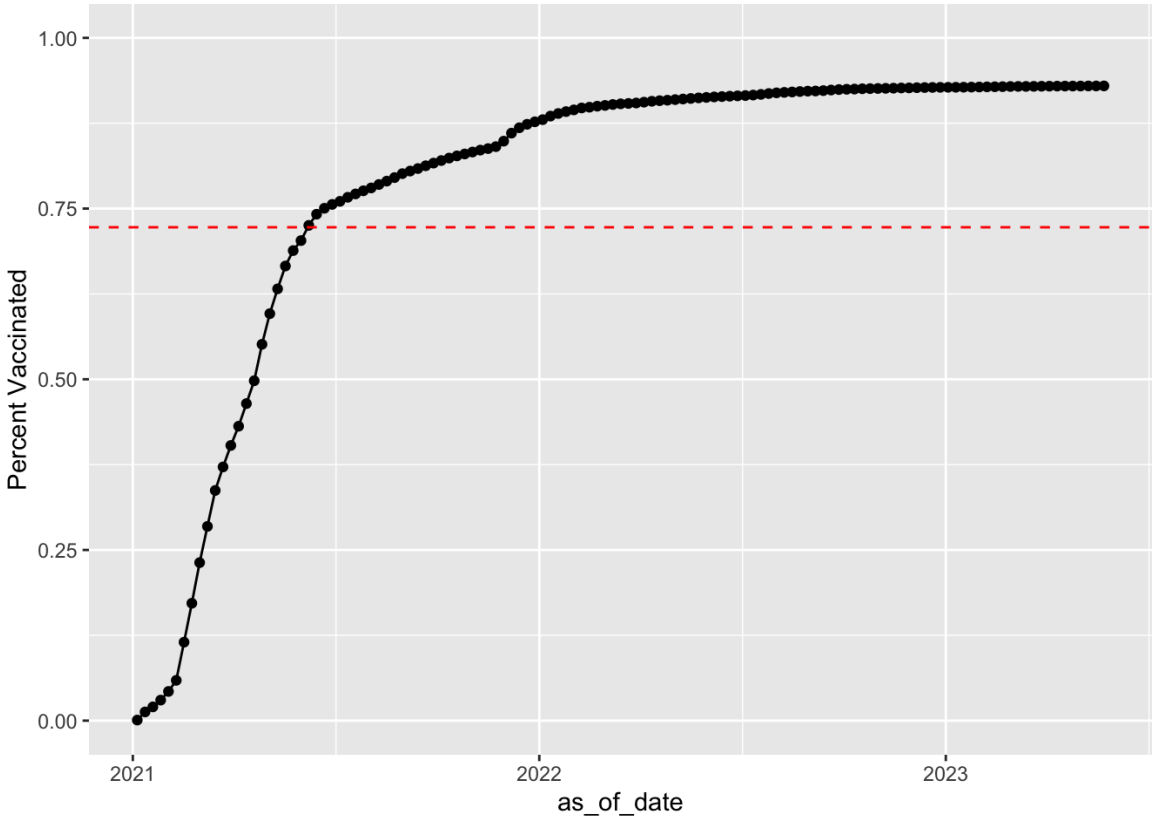
```
# Subset to all CA areas with a population as large as 92037
vax.36 <- filter(vax, age5_plus_population > 36144 &
                 as_of_date == "2023-05-23")
```

```
mean_percent <- mean(vax.36$percent_of_population_fully_vaccinated)
mean_percent *100
```

```
## [1] 72.25892
```

##It is 72.26%

```
#Now add it to the graph
library(ggplot2)
ggplot(ucsd) +
  aes(x=as_of_date,
      y=percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1)+ geom_hline(aes(yintercept=mean_percent), color="red", linetype="dashed") +
  ylim(c(0,1)) +
  labs("Date", y="Percent Vaccinated")
```



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 “2023-05-23”?

```
#when summary comes into play...
summary(vax.36$percent_of_population_fully_vaccinated)
```

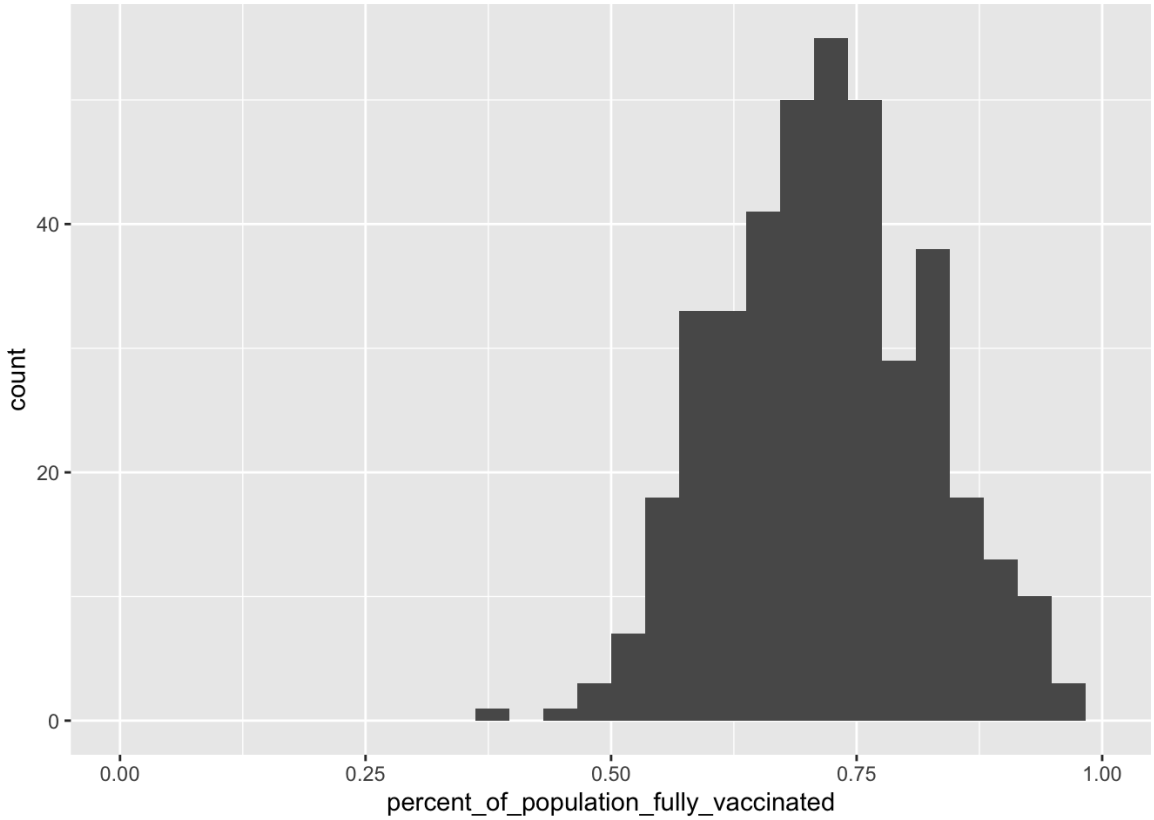
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3816  0.6469  0.7207  0.7226  0.7924  1.0000
```

Q18. Using ggplot generate a histogram of this data

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

```
## `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?

```
#percent for 92109
library(dplyr)
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 <dbl>
0.675129

1 row

```
#percent for 92040
library(dplyr)
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>
0.533991

1 row

##Based on these two percentages above, both of them are bewlow the average value I calculated previously.

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(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 rate across California",
       subtitle="only areas with a population above 36k are shown") +
  geom_hline(yintercept =mean_percent , linetype="dashed")
```

```
## Warning: Removed 185 rows containing missing values (`geom_line()`).
```