Class 14 Mini-project

Samuel Do (PID:A15803613)

3/3/2022

```
# Import 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
##                                                        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?
# persons_fully_vaccinated
#[Q2] What column details the Zip code tabulation area?
# zip_code_tabulation_area
list(unique(vax$as_of_date, TRUE))
```

```
## [[1]]
##  [1] "2021-01-05" "2021-01-12" "2021-01-19" "2021-01-26" "2021-02-02"
##  [6] "2021-02-09" "2021-02-16" "2021-02-23" "2021-03-02" "2021-03-09"
## [11] "2021-03-16" "2021-03-23" "2021-03-30" "2021-04-06" "2021-04-13"
## [16] "2021-04-20" "2021-04-27" "2021-05-04" "2021-05-11" "2021-05-18"
## [21] "2021-05-25" "2021-06-01" "2021-06-08" "2021-06-15" "2021-06-22"
## [26] "2021-06-29" "2021-07-06" "2021-07-13" "2021-07-20" "2021-07-27"
## [31] "2021-08-03" "2021-08-10" "2021-08-17" "2021-08-24" "2021-08-31"
## [36] "2021-09-07" "2021-09-14" "2021-09-21" "2021-09-28" "2021-10-05"
## [41] "2021-10-12" "2021-10-19" "2021-10-26" "2021-11-02" "2021-11-09"
## [46] "2021-11-16" "2021-11-23" "2021-11-30" "2021-12-07" "2021-12-14"
## [51] "2021-12-21" "2021-12-28" "2022-01-04" "2022-01-11" "2022-01-18"
## [56] "2022-01-25" "2022-02-01" "2022-02-08" "2022-02-15" "2022-02-22"
## [61] "2022-03-01"
```

```
#[Q3] What is the earliest date in this dataset?
# 2021-01-05 is the earliest date in the dataset.
#[Q4] What is the latest date in this dataset?
# 2022-03-01 is the latest date in this dataset.
```

```
#install.packages("skimr")
library(skimr)
skimr::skim(vax)
```

Data summary

Name	vax
Number of rows	107604
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	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	▃▆▇▇

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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 9 numeric columns if the zip_code_tabulation_area column is not included.
# [Q6] Note that there are "missing values" in the dataset. How many NA values there in the persons_fully_vaccinated column?
# There are 18338 NA values in the person_fully_vaccinated column.
# [Q7] What percent of persons_fully_vaccinated values are missing (to 2 significant figures)?
# 17% of persons_fully_vaccinated values are missing.
# [Q8] [Optional]: Why might this data be missing?
# This data might be missing due to the lack of data on vaccination statuses, which requires either voluntary responses or l
egal permission to collect.
```

```
library(lubridate)
```

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

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

```
today()
```

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

```
# today() - vax$as_of_date[1] (ERROR!)
# Must specify use of year-month-day format first
vax$as_of_date <- ymd(vax$as_of_date)
today() - vax$as_of_date[1]
```

```
## Time difference of 426 days
```

```
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?
# 6 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, TRUE)))
```

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

```
# There are 61 unique dates in the dataset.
```

```
#install.packages("zipcodeR")
library(zipcodeR)
geocode_zip('92037')
```

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

```
# Calculate the distance between the centroids of any two ZIP codes in miles
zip_distance('92037','92109')
```

```
##   zipcode_a zipcode_b distance
## 1     92037     92109     2.33
```

```
# Pull census data about 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>
```

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

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] 6527
```

```
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, TRUE))
```

```
## [1] 107
```

```
# There are 107 distinct zip codes listed for 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 San Diego county zip code with the largest 12+ population is 92154.
```

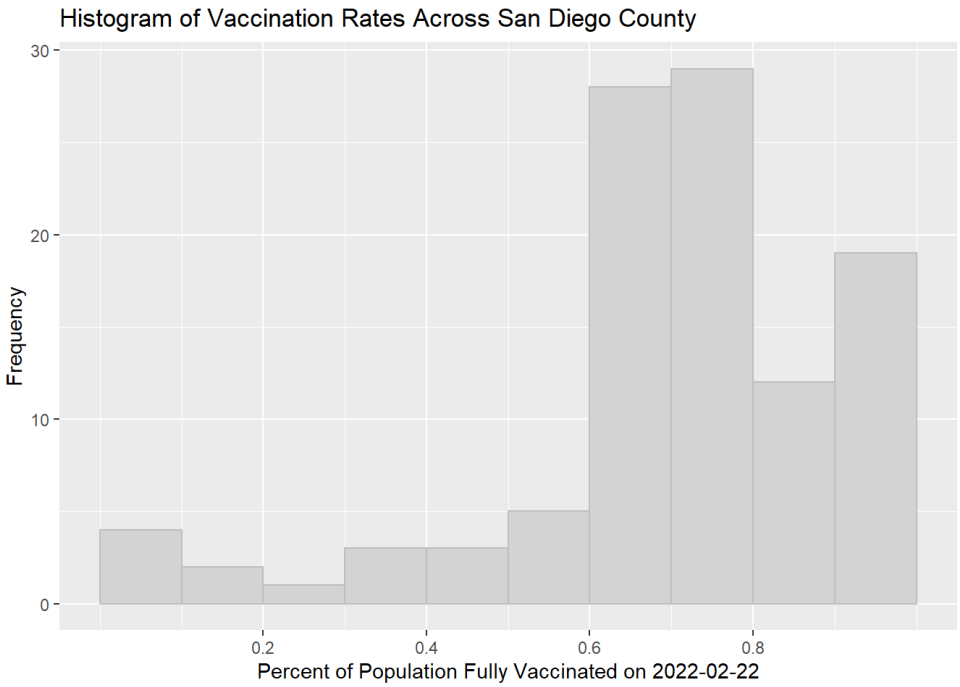
```
# Using dplyr select all San Diego "county" entries on "as_of_date" "2022-02-22"
sd2 <- filter(sd, as_of_date =="2022-02-22")
#[Q13] What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2022-02-22"?
mean(sd2$percent_of_population_fully_vaccinated, na.rm=TRUE)
```

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

```
# 70.42% of all San Diego County were fully vaccinated as of 2022-02-22.
#[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(sd2, aes(x=percent_of_population_fully_vaccinated)) +
  geom_histogram(color="gray", fill="lightgray", binwidth=0.1, origin=0) +
  labs(title="Histogram of Vaccination Rates Across San Diego County") +
  xlab("Percent of Population Fully Vaccinated on 2022-02-22") +
  ylab("Frequency") +
  scale_x_continuous(breaks=c(0.2,0.4,0.6,0.8))
```

```
## Warning: `origin` is deprecated. Please use `boundary` instead.
```

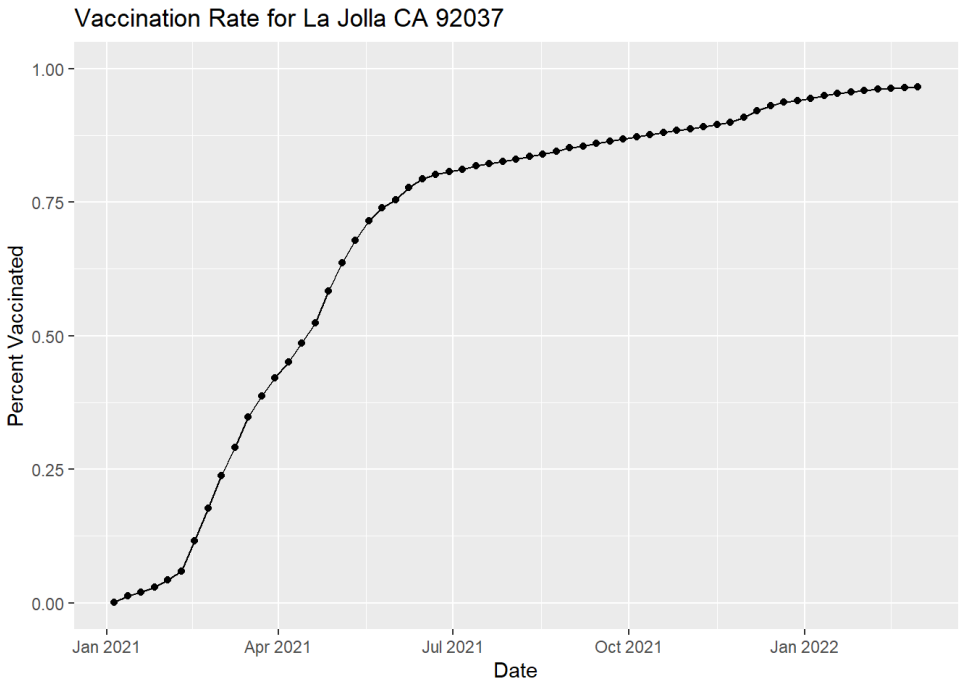
```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```




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

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

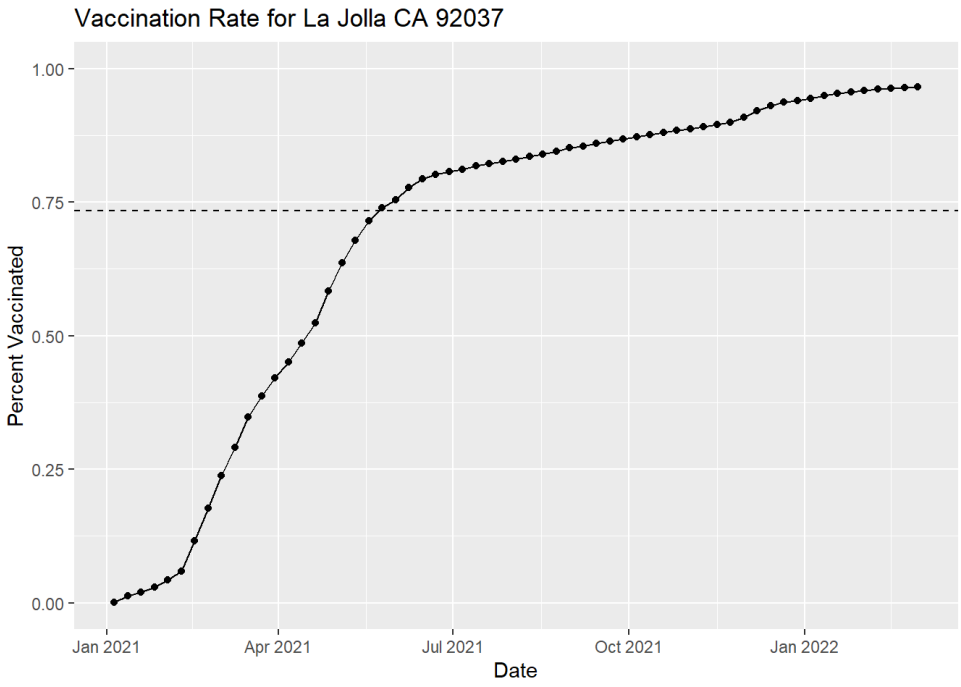
```
#[Q15] Using ggplot make a graph of the vaccination rate time course for the 92037 ZIP code area
ggplot(ucsd) +
  aes(x=as_of_date,
      y=percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) +
  ylim(c(0,1)) +
  labs(title="Vaccination Rate for La Jolla CA 92037", x="Date", y="Percent Vaccinated")
```




```
# 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)
#[Q16] Calculate the mean "Percent of Population Fully Vaccinated" for ZIP code areas with a population as large as 92037 (L
a Jolla) as_of_date "2022-02-22". Add this as a straight horizontal line to your plot from above with the geom_hline() funct
ion?
mean(vax.36$percent_of_population_fully_vaccinated)
```

```
## [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(yintercept = mean(vax.36$percent_of_population_fully_vaccinated), linetype='dashed')
```



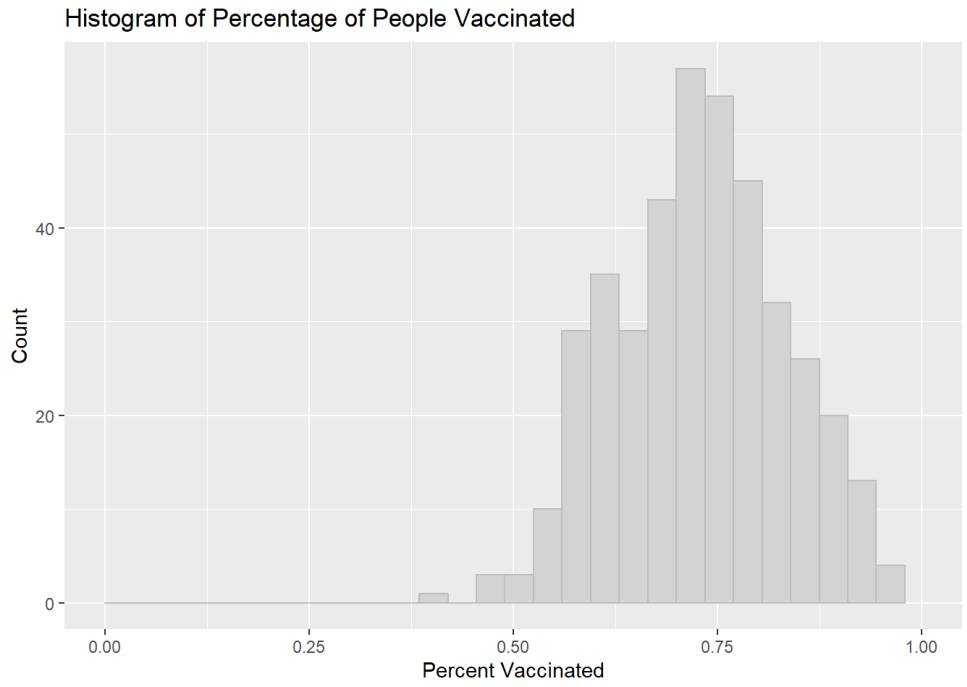

```
#[Q17] What is the 6 number summary (Min, 1st Qu., Median, Mean, 3rd Qu., and Max) of the "Percent of Population Fully Vacci
nated" values for ZIP code areas with a population as large as 92037 (La Jolla) as_of_date "2022-02-22"?
summary(vax.36$percent_of_population_fully_vaccinated)
```

```
##    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(color="gray", fill="lightgray", binwidth = 0.035, origin=0) +
  labs(title="Histogram of Percentage of People Vaccinated") +
  xlab("Percent Vaccinated") +
  ylab("Count") +
  xlim(0,1)
```

```
## Warning: `origin` is deprecated. Please use `boundary` instead.
```

```
## Warning: Removed 1 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?
vax.1 <- vax %>% filter(as_of_date == "2022-02-22") %>%
  filter(zip_code_tabulation_area=="92040")
vax.2 <- vax %>% filter(as_of_date == "2022-02-22") %>%
  filter(zip_code_tabulation_area=="92109")
mean(vax.1$percent_of_population_fully_vaccinated) > mean(vax.36$percent_of_population_fully_vaccinated)
```

```
## [1] FALSE
```

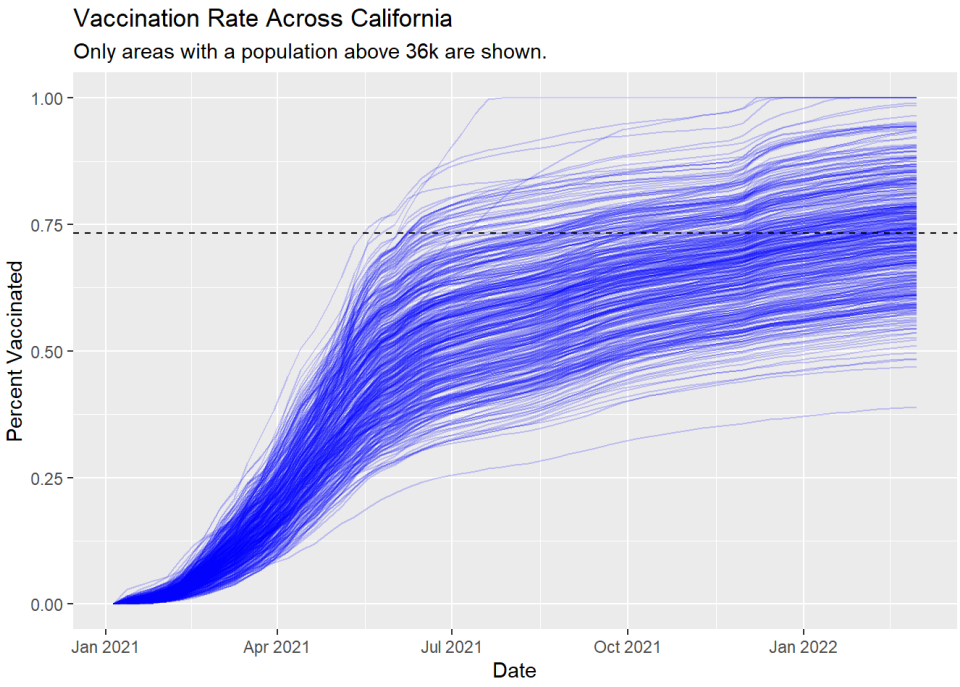
```
mean(vax.2$percent_of_population_fully_vaccinated) > mean(vax.36$percent_of_population_fully_vaccinated)
```

```
## [1] FALSE
```

```
# Both 92109 and 92040 ZIP code areas are below the average value of percent_of_population_fully_vaccinated as of 2022-02-2
2.

#[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(vax.36$percent_of_population_fully_vaccinated), 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 unfortunately will not be traveling for Spring Break for personal reasons, but I am excited for in-person classes in the future!
```