class14R

Katelyn Brown A15891811

3/3/2022

Read our vaccination data.

Downloaded the most recently dated “Statewide COVID-19 Vaccines Administered by ZIP Code” CSV.

```r
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?

The persons_fully_vaccinated column details the total number of people fully vaccinated.

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

The zip_code_tabulation_area column details the Zip code tabulation area.

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

The earliest date in the dataset is 2021-01-05.

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

```r
vax$as_of_date[nrow(vax)]
```

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

The latest date in the dataset is 2022-03-01.

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

Table 1: Data summary

Name	vax
Number of rows	107604
Number of columns	15
Column type frequency:	
character	5

Table 1: Data summary

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	
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 in the dataset.

> 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 persons_fully_vaccinated column.

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

17% of persons_fully_vaccinated is missing.

> Q8. Why might this data be missing?

This data could be missing due to lack of reporting, or possibly lack of access. Some counties may also not be reporting vaccination rates.

```r
# install.packages("lubridate")
library(lubridate)
```

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

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

```r
today()
```

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

```r
age <- today() - ymd("2001-11-16")
age
```

```
## Time difference of 7412 days
```

```r
time_length(age, "year")
```

```
## [1] 20.29295
```

```r
vax$as_of_date <- ymd(vax$as_of_date)
today() - vax$as_of_date[1]
```

```
## Time difference of 422 days
```

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

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

```
## Time difference of 2 days
```

```r
# Determine how many days the dataset spans
vax$as_of_date[nrow(vax)] - vax$as_of_date[1]
```

```
## Time difference of 420 days
```

2 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)?

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

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

There are 61 unique dates in the dataset.

Working with Zip Codes

First, download the zipcodeR package and load it in the library.

```
# install.packages("zipcodeR")
library(zipcodeR)
```

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

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

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

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

```
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>
```

```
zipdata <- reverse_zipcode(vax$zip_code_tabulation_area)
zipdata
```

```
## # A tibble: 1,764 x 24
##    zipcode zipcode_type major_city  post_office_city common_city_list county
##    <chr>   <chr>        <chr>       <chr>                      <blob> <chr>
##  1 90001   Standard     Los Angeles Los Angeles, CA        <raw 44 B> Los Angel~
##  2 90002   Standard     Los Angeles Los Angeles, CA        <raw 47 B> Los Angel~
##  3 90003   Standard     Los Angeles Los Angeles, CA        <raw 23 B> Los Angel~
##  4 90004   Standard     Los Angeles Los Angeles, CA        <raw 34 B> Los Angel~
##  5 90005   Standard     Los Angeles Los Angeles, CA        <raw 34 B> Los Angel~
##  6 90006   Standard     Los Angeles Los Angeles, CA        <raw 23 B> Los Angel~
##  7 90007   Standard     Los Angeles Los Angeles, CA        <raw 37 B> Los Angel~
##  8 90008   Standard     Los Angeles Los Angeles, CA        <raw 53 B> Los Angel~
##  9 90010   Standard     Los Angeles Los Angeles, CA        <raw 23 B> Los Angel~
## 10 90011   Standard     Los Angeles Los Angeles, CA        <raw 23 B> Los Angel~
## # ... with 1,754 more rows, and 18 more variables: state <chr>, 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>
```

Filter to only work with San Diego county.

```
# Base R
dim(vax[vax$county == "San Diego", ])
```

```
## [1] 6527   15
```

```
# Dplyr
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")
dim(sd)
```

```
## [1] 6527   15
```

> 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 in San Diego County.

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

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

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

92154 has the largest 12+ population in the dataset.

> Q13. What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2022-03-01"?

```
sd$as_of_date[nrow(sd)]
```

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

```
# Filter to day
sd.latest <- filter(sd, as_of_date == "2022-03-01")
mean(sd.latest$percent_of_population_fully_vaccinated, na.rm = TRUE)*100
```

```
## [1] 70.52904
```

```
summary(sd.latest$percent_of_population_fully_vaccinated, na.rm = T)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
## 0.01017 0.65132 0.72452 0.70529 0.82567 1.00000       1
```

70.53% of San Diego county is fully vaccinated as of 2022-03-01.

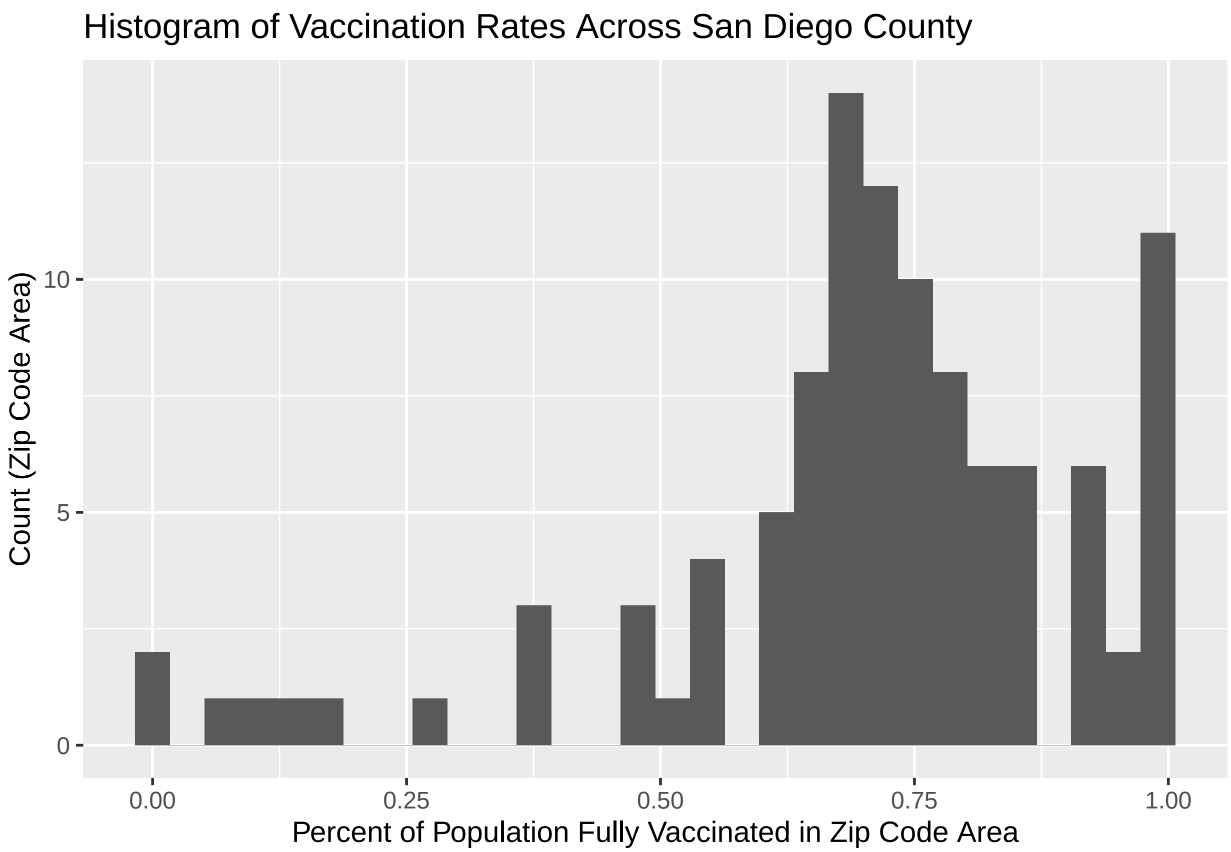
> Q14. Using either ggplot or base R graphics make a summary figure that shows the distribution of Percent of Population Fully Vaccinated values as of “2022-03-01”?

```
library(ggplot2)
ggplot(sd.latest) + aes(x = sd.latest$percent_of_population_fully_vaccinated) + geom_histogram() + labs
```

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

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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



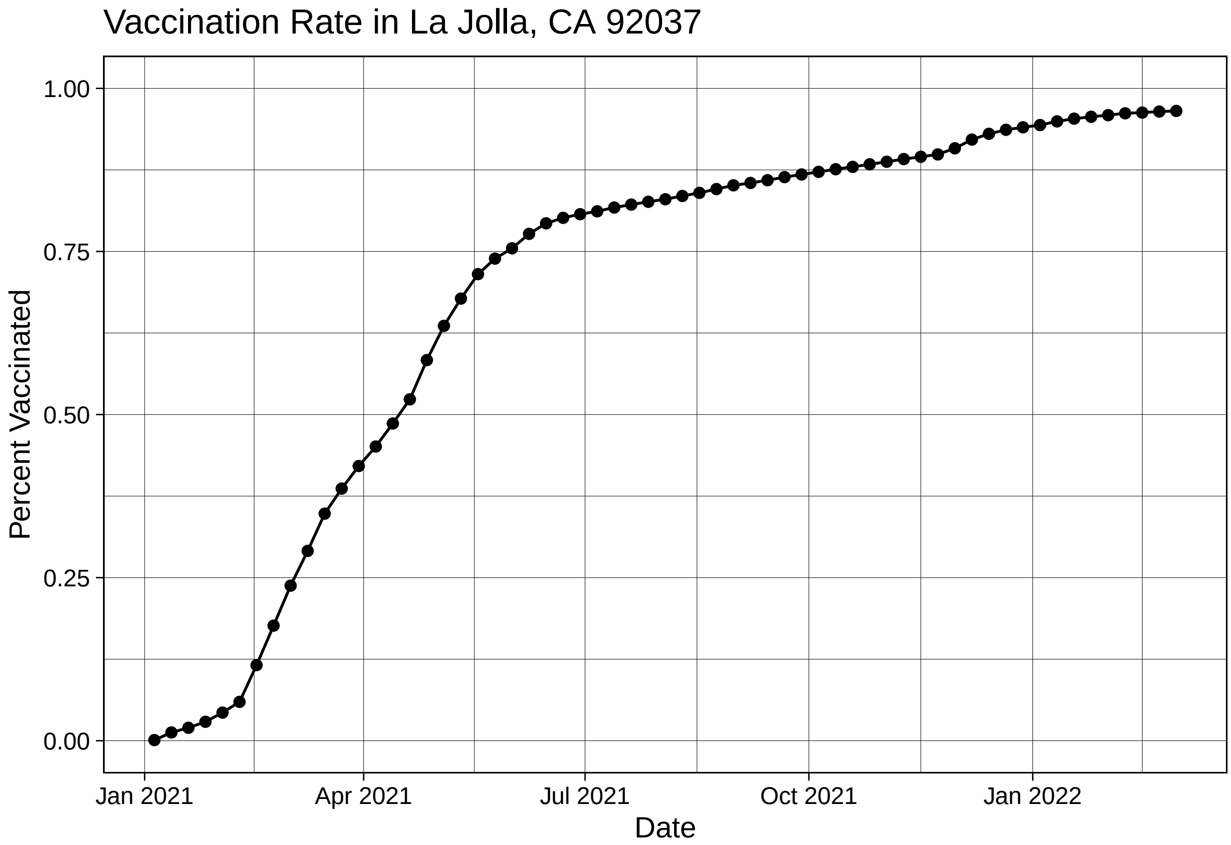
Filter to focus on UCSD.

```
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.

```
baseplot <- ggplot(ucsd) +
  aes(x = as_of_date, y = percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) + theme_linedraw() +
  ylim(c(0,1)) +
  labs(x = "Date", y="Percent Vaccinated", title = "Vaccination Rate in La Jolla, CA 92037")
baseplot
```



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 "2022-03-01". Add this as a straight horizontal line to your plot from above with the geom_hline() function?

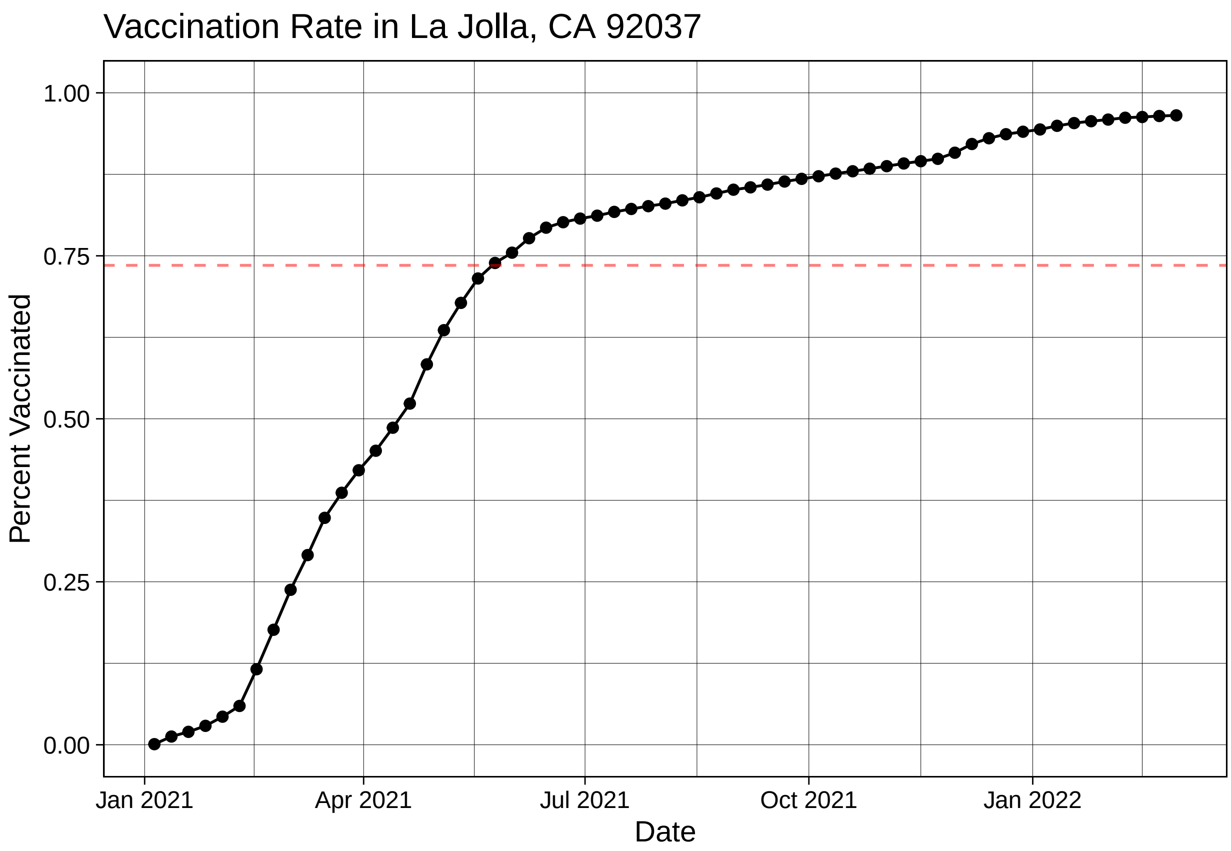
```
# Subset other CA zip codes with populations as big as 92037
vax.36 <- filter(vax, age5_plus_population > 36144 & as_of_date == "2022-03-01")
head(vax.36)
```

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction      county
## 1 2022-03-01                    95628                Sacramento  Sacramento
## 2 2022-03-01                    90808                Long Beach Los Angeles
## 3 2022-03-01                    92507                 Riverside   Riverside
## 4 2022-03-01                    92626                    Orange      Orange
## 5 2022-03-01                    93257                    Tulare      Tulare
## 6 2022-03-01                    90011               Los Angeles Los Angeles
##   vaccine_equity_metric_quartile                 vem_source
## 1                              3 Healthy Places Index Score
## 2                              4 Healthy Places Index Score
## 3                              1 Healthy Places Index Score
## 4                              3 Healthy Places Index Score
## 5                              1 Healthy Places Index Score
## 6                              1 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1               35579.0                38694                    28842
## 2               33952.3                37179                    29383
## 3                51432.5                55253                    34455
## 4                44238.8                47883                    33767
## 5                61519.8                70784                    42919
## 6                87902.8               101902                    65342
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                         1990                               0.745387
## 2                         2112                               0.790312
## 3                         3947                               0.623586
## 4                         2937                               0.705198
## 5                         5868                               0.606338
## 6                        15255                               0.641224
##   percent_of_population_partially_vaccinated
## 1                                   0.051429
## 2                                   0.056806
## 3                                   0.071435
## 4                                   0.061337
## 5                                   0.082900
## 6                                   0.149703
##   percent_of_population_with_1_plus_dose booster_recip_count redacted
## 1                               0.796816               16913       No
## 2                               0.847118               17253       No
## 3                               0.695021               15073       No
## 4                               0.766535               17595       No
## 5                               0.689238               17740       No
## 6                               0.790927               19928       No
```

```
# Find mean of this data
vax.36mean <- mean(vax.36$percent_of_population_fully_vaccinated, na.rm = TRUE)
vax.36mean
```

```
## [1] 0.7353974
```

```
# Add mean to baseplot
baseplot + geom_hline(yintercept = vax.36mean, linetype = 2, alpha = 0.5, color = "red")
```



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 "2022-03-01"?

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

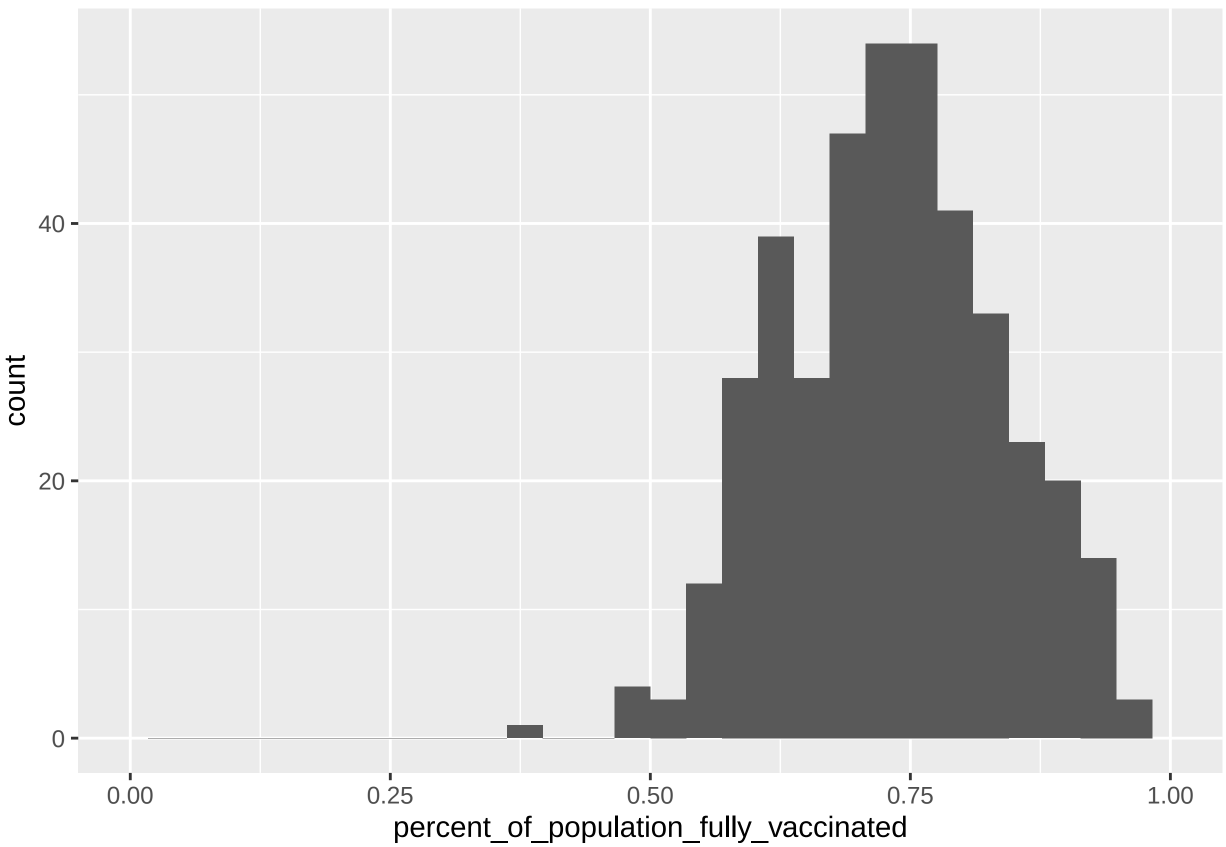
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3890  0.6554  0.7350  0.7354  0.8044  1.0000
```

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?

```
vax %>% filter(as_of_date == "2022-03-01") %>%
  filter(zip_code_tabulation_area=="92109") %>%
  select(percent_of_population_fully_vaccinated)
```

```
##   percent_of_population_fully_vaccinated
## 1                               0.723778
```

```
vax %>% filter(as_of_date == "2022-03-01") %>%
  filter(zip_code_tabulation_area=="92040") %>%
  select(percent_of_population_fully_vaccinated)
```

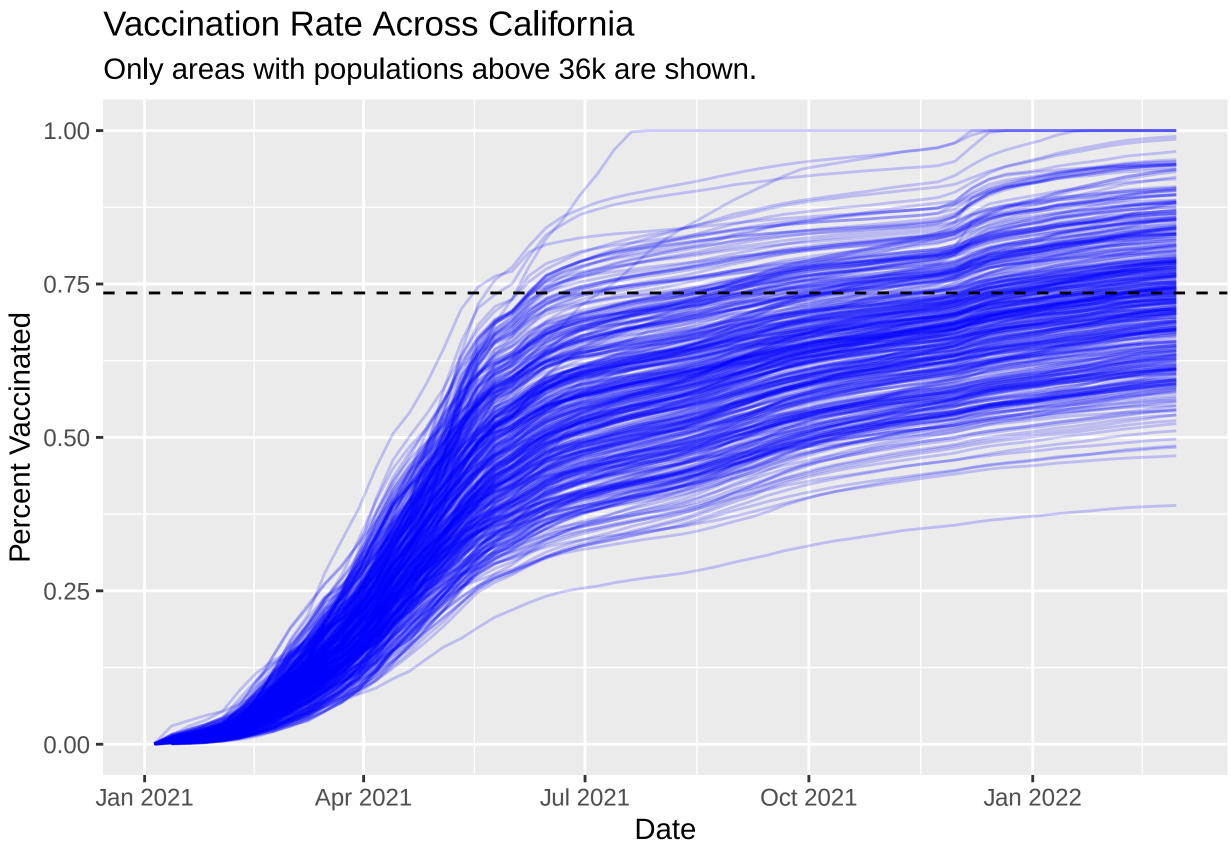
```
##   percent_of_population_fully_vaccinated
## 1                               0.551981
```

92109 and 92040 zip codes are below the average of 0.735 that was 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.

```r
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 populations above 36k are shown.") +
  geom_hline(yintercept = vax.36mean, linetype = 2)
```

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


Vaccination Rate Across California
Only areas with populations above 36k are shown.
Percent Vaccinated
Jan 2021
Apr 2021
Jul 2021
Oct 2021
Jan 2022
Date