Lab17

Jordan Laxa

```
vax <- read.csv("covid19vaccinesbyzipcode_test.csv")
head(vax)
```

```
  as_of_date zip_code_tabulation_area local_health_jurisdiction          county
1 2021-01-05                   93609                     Fresno          Fresno
2 2021-01-05                   94086                Santa Clara     Santa Clara
3 2021-01-05                   94304                Santa Clara     Santa Clara
4 2021-01-05                   94110              San Francisco   San Francisco
5 2021-01-05                   93420            San Luis Obispo San Luis Obispo
6 2021-01-05                   93454              Santa Barbara   Santa Barbara
  vaccine_equity_metric_quartile                vem_source
1                              1 Healthy Places Index Score
2                              4 Healthy Places Index Score
3                              4 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 tot_population
1                4396.3                 4839           5177
2               42696.0                46412          50477
3                3263.5                 3576           3852
4               64350.7                68320          72380
5               26694.9                29253          30740
6               32043.4                36446          40432
  persons_fully_vaccinated persons_partially_vaccinated
1                       NA                           NA
2                       11                          640
3                       NA                           NA
4                       18                         1262
5                       NA                           NA
6                       NA                           NA
```

```
  percent_of_population_fully_vaccinated
1                                     NA
2                                0.000218
3                                     NA
4                                0.000249
5                                     NA
6                                     NA
  percent_of_population_partially_vaccinated
1                                         NA
2                                    0.012679
3                                         NA
4                                    0.017436
5                                         NA
6                                         NA
  percent_of_population_with_1_plus_dose booster_recip_count
1                                     NA                  NA
2                                0.012897                 NA
3                                     NA                  NA
4                                0.017685                 NA
5                                     NA                  NA
6                                     NA                  NA
  bivalent_dose_recip_count eligible_recipient_count
1                        NA                       1
2                        NA                      11
3                        NA                       6
4                        NA                      18
5                        NA                       4
6                        NA                       5
                                                                    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?

Q1A. persons_fully_vaccinated

Q2. What column details the Zip code tabulation area?

Q2A. zip_code_tabulation_area

Q3. What is the earliest date in this dataset?

Q3A. 2021-01-05

```
min(vax[,"as_of_date"])
```

```
[1] "2021-01-05"
```

Q4. What is the latest date in this dataset?

Q4A. 2023-03-07

```
max(vax[,"as_of_date"])
```

```
[1] "2023-03-07"
```

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

Table 1: Data summary

Name	vax
Number of rows	201096
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	114	0
local_health_jurisdiction	0	1	0	15	570	62	0
county	0	1	0	15	570	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	17.38	90001	92257.75	93658.50	95380.50	97635.0	
vaccine_equity_metric_quartile	9918	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	
age12_plus_population	0	1.00	18895.01	8993.87	0	13346.95	13685.13	21756.12	28556.7	
age5_plus_population	0	1.00	20875.22	11105.97	0	14605.50	15364.00	24877.00	31902.0	
tot_population	9804	0.95	23372.72	12628.50	12	21260.00	18714.00	28168.00	111165.0	
persons_fully_vaccinated	16621	0.92	13990.31	5073.66	11	9932.00	8589.00	23346.00	37575.0	
persons_partially_vaccinated	16621	0.92	1702.31	2033.32	11	1650.00	1197.00	2536.00	39973.0	
percent_of_population_fully_vaccinated	20965	0.90	0.57	0.25	0	0.42	0.61	0.74	1.0	
percent_of_population_partially_vaccinated	20965	0.90	0.08	0.09	0	0.05	0.06	0.08	1.0	
percent_of_population_with_1_plus_dose	22009	0.89	0.63	0.24	0	0.49	0.67	0.81	1.0	
booster_recip_count	72997	0.64	5882.76	7219.00	11	300.00	2773.00	9510.00	59593.0	
bivalent_dose_recip_count	158776	0.21	2978.23	3633.03	11	193.00	1467.50	4730.25	27694.0	
eligible_recipient_count	0	1.00	12830.83	4928.64	0	8507.00	6369.00	22014.00	37248.0	

Q5. How many numeric columns are in this dataset?

Q5A. 13

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

```
nrow(vax[vax$"persons_fully_vaccinated" == "NA",])
```

```
[1] 16621
```

Q6A. 16621

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

```
nrow(vax[vax$"persons_fully_vaccinated" == "NA",]) / nrow(vax[vax$"persons_fully_vaccinate
```

```
[1] 8.265207
```

Q7A. 8.27%

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

Q8A. It might be missing because some clinics might have no reported their vaccination results yet.

```
  library(lubridate)
```

```
Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

```
  today()
```

```
[1] "2023-03-13"
```

```
  vax$as_of_date <- ymd(vax$as_of_date)
```

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

```
Time difference of 797 days
```

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

```
Time difference of 791 days
```

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

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

```
Time difference of 6 days
```

Q9A. 6 days

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

```
  length(unique(vax[vax$as_of_date, ]))
```

```
[1] 18
```

Q10A. There are 18 unique dates.

```r
library(zipcodeR)
```

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

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

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

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

```r
reverse_zipcode(c('92037', "92109") )
```

```
# A tibble: 2 x 24
  zipcode zipcode_~1 major~2 post_~3 common_c~4 county state   lat   lng timez~5
  <chr>   <chr>      <chr>   <chr>       <blob> <chr>  <chr> <dbl> <dbl> <chr>
1 92037   Standard   La Jol~ La Jol~ <raw 20 B> San D~ CA     32.8 -117. Pacific
2 92109   Standard   San Di~ San Di~ <raw 21 B> San D~ CA     32.8 -117. Pacific
# ... with 14 more variables: 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>, and abbreviated variable names
#   1: zipcode_type, 2: major_city, 3: post_office_city, ...
```

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

```r
sd <- vax[vax$county == "San Diego" , ]
nrow(sd)
```

```
[1] 12198
```

```
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" &
                as_of_date > 2022-11-15)

nrow(sd)
```

```
[1] 12198
```

```
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(filter(vax, county == "San Diego")))
```

```
[1] 18
```

Q11A. 18

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

```
sd[which.max(sd$county == "San Diego"),]
```

```
  as_of_date zip_code_tabulation_area local_health_jurisdiction    county
1 2021-01-05                   91911                San Diego San Diego
  vaccine_equity_metric_quartile               vem_source
1                              2 Healthy Places Index Score
  age12_plus_population age5_plus_population tot_population
1               71642.8                79225          84026
  persons_fully_vaccinated persons_partially_vaccinated
1                       28                         1420
  percent_of_population_fully_vaccinated
1                               0.000333
  percent_of_population_partially_vaccinated
1                                     0.0169
  percent_of_population_with_1_plus_dose booster_recip_count
1                               0.017233                  NA
  bivalent_dose_recip_count eligible_recipient_count
1                        NA                       28
                                                                redacted
1 Information redacted in accordance with CA state privacy requirements
```

Q12A. 91911 has the largest 12+ population.

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

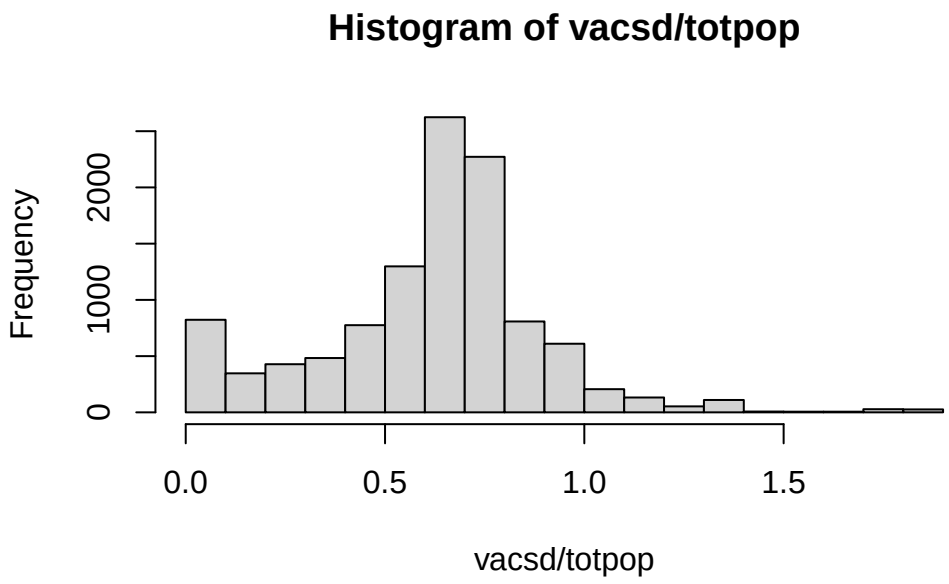
```
fullsd <- sum(sd$"persons_fully_vaccinated", na.rm = TRUE)
totsd <- sum(sd$"tot_population", na.rm = TRUE)
fullsd / totsd * 100
```

```
[1] 60.45631
```

Q13A. 60.46%

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 "2022-11-15"?

```
vacsd <- (sd$"persons_fully_vaccinated")
totpop <- (sd$"tot_population")
hist(vacsd/totpop)
```

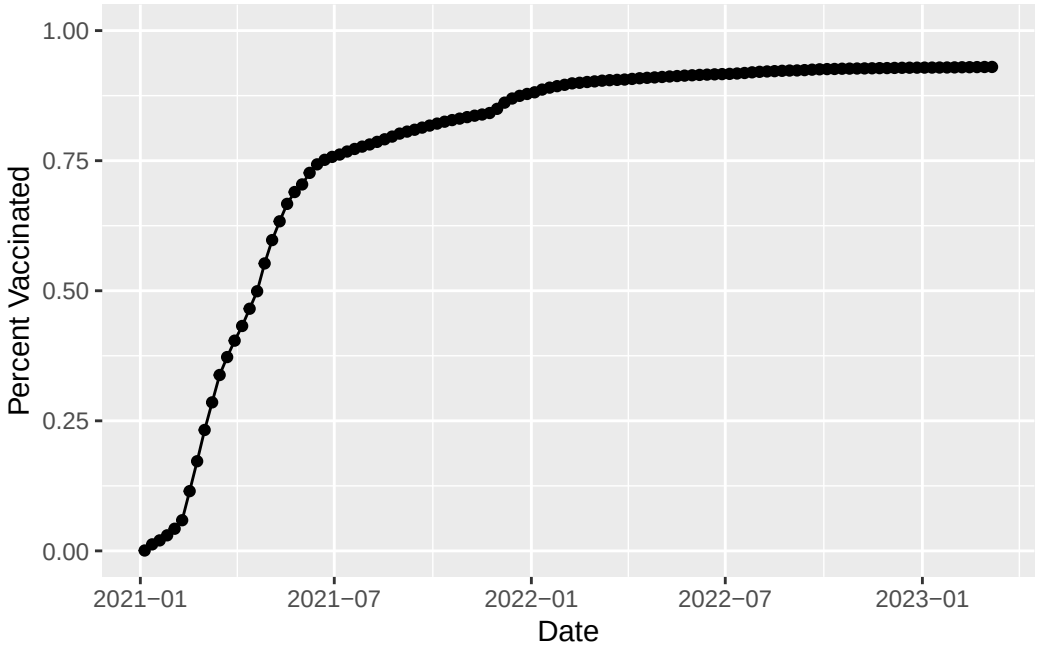



```
ucsd <- filter(sd, zip_code_tabulation_area=="92037")
ucsd[1,]$age5_plus_population
```

```
[1] 36144
```

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

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




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

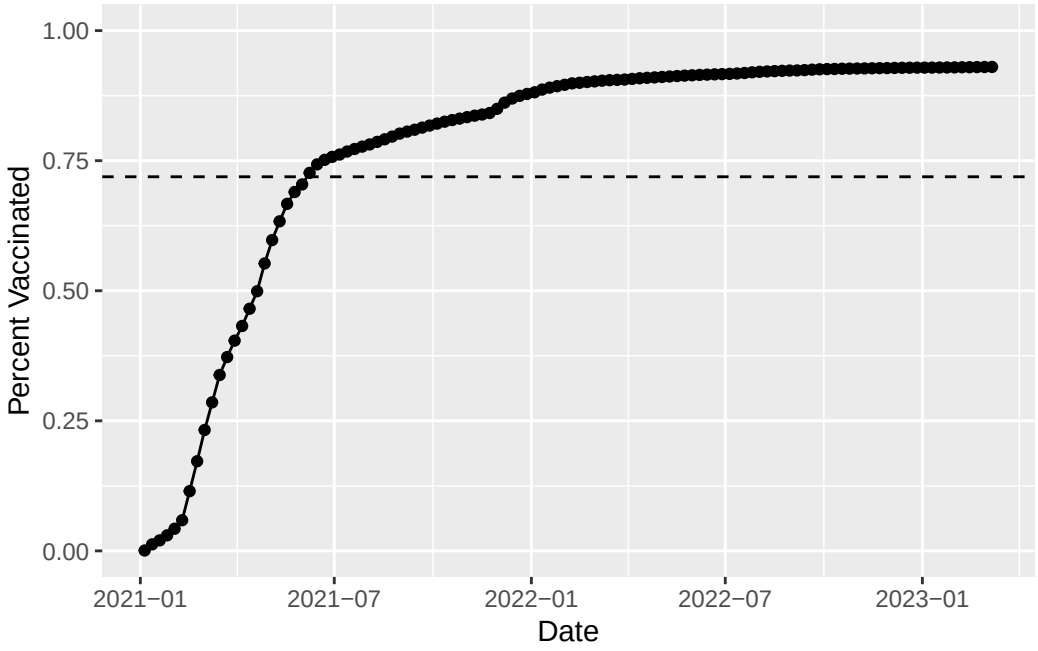
Q16. Calculate the mean "Percent of Population Fully Vaccinated" for ZIP code areas with a population as large as 92037 (La Jolla) as_of_date "2022-11-15". Add this as a straight horizontal line to your plot from above with the geom_hline() function?

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

```
[1] 0.7190967
```

Q16A.

```
plot + geom_hline(yintercept=0.7190967, linetype = "dashed")
```



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"?

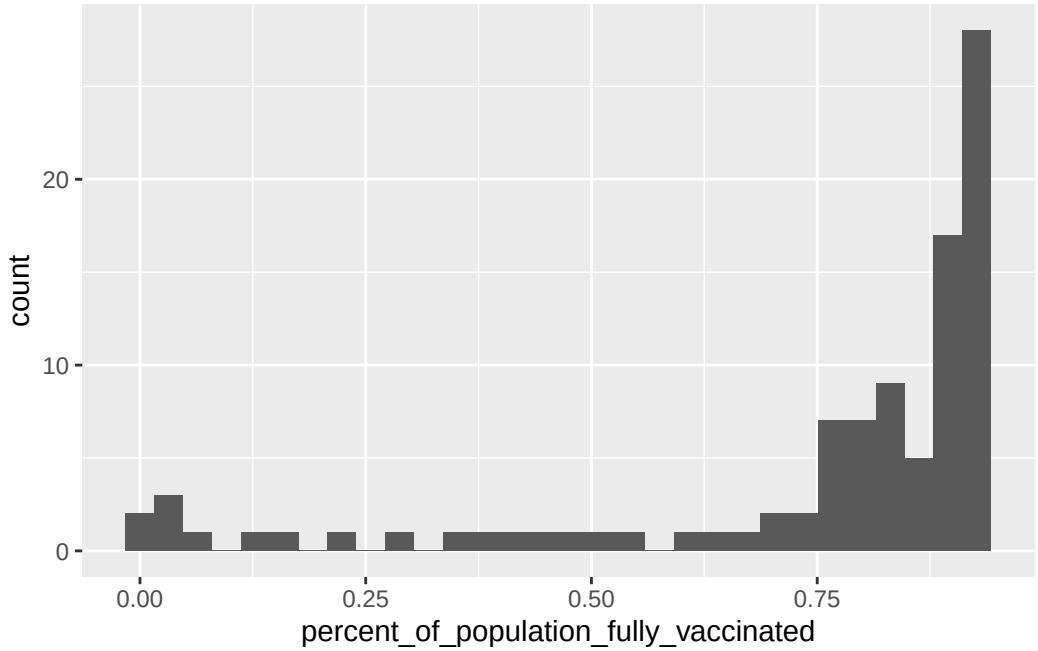
Q17A.

```
ucsdas <- filter(ucsd, as_of_date < "2022-11-15")
summary(ucsdas$percent_of_population_fully_vaccinated)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00076 0.75165 0.86153 0.74895 0.91349 0.92737
```

```
ggplot(ucsdas) + aes(percent_of_population_fully_vaccinated) + geom_histogram(aes(y=..coun
```

```
Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
i Please use `after_stat(count)` instead.
```



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                               0.548979
```

```
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                               0.692832
```

Q19A. They are below the average value of 75%.

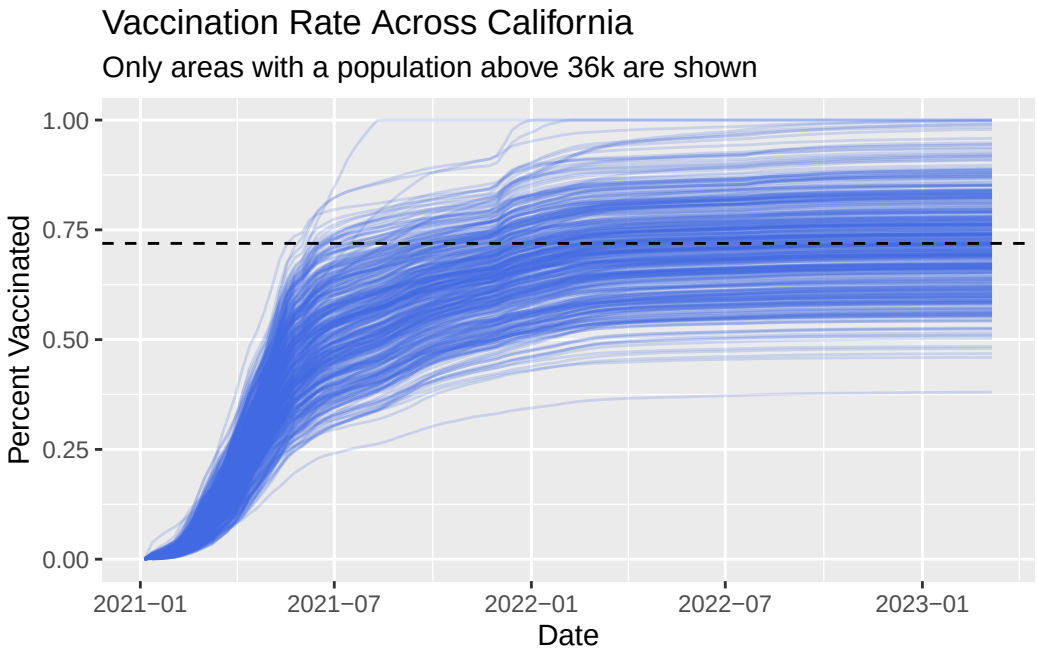
Q20. Finally make a time course plot of vaccination progress for all areas in the full dataset with a age5_plus_population > 36144.

Q20A.

```
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="royalblue") +
  ylim(c(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 = 0.7190967, linetype="dashed")
```

```
Warning: Removed 183 rows containing missing values (`geom_line()`).
```



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

Q21A. I would love to come back for meeting in class, however the quarter would be over by the time break ends :D.