COVID-19 Vaccination Rates Mini-Project

Kyle

We will be examining and comparing the Covid-19 vaccination rates in San Diego.

Start by downloading the most recently dated "Statewide COVID-19 Vaccines Administered by ZIP Code" CSV file.

```r
# Import vaccination data
vax <- read.csv("https://data.chhs.ca.gov/dataset/ead44d40-fd63-4f9f-950a-3b0111074de8/res
head(vax)
```

```
  as_of_date zip_code_tabulation_area local_health_jurisdiction          county
1 2021-01-05                    92240                 Riverside       Riverside
2 2021-01-05                    91302               Los Angeles     Los Angeles
3 2021-01-05                    93420           San Luis Obispo San Luis Obispo
4 2021-01-05                    91901                 San Diego       San Diego
5 2021-01-05                    94110             San Francisco   San Francisco
6 2021-01-05                    91902                 San Diego       San Diego
  vaccine_equity_metric_quartile                 vem_source
1                              1 Healthy Places Index Score
2                              4 Healthy Places Index Score
3                              3 Healthy Places Index Score
4                              3 Healthy Places Index Score
5                              4 Healthy Places Index Score
6                              4 Healthy Places Index Score
  age12_plus_population age5_plus_population tot_population
1               29270.5                33093          35278
2               23163.9                25899          26712
3               26694.9                29253          30740
4               15549.8                16905          18162
5               64350.7                68320          72380
6               16620.7                18026          18896
  persons_fully_vaccinated persons_partially_vaccinated
1                       NA                           NA
```

```
2                       15                          614
3                       NA                           NA
4                       NA                           NA
5                       17                         1268
6                       15                          397
  percent_of_population_fully_vaccinated
1                                     NA
2                               0.000562
3                                     NA
4                                     NA
5                               0.000235
6                               0.000794
  percent_of_population_partially_vaccinated
1                                         NA
2                                   0.022986
3                                         NA
4                                         NA
5                                   0.017519
6                                   0.021010
  percent_of_population_with_1_plus_dose booster_recip_count
1                                     NA                  NA
2                               0.023548                  NA
3                                     NA                  NA
4                                     NA                  NA
5                               0.017754                  NA
6                               0.021804                  NA
  bivalent_dose_recip_count eligible_recipient_count
1                        NA                        2
2                        NA                       15
3                        NA                        4
4                        NA                        8
5                        NA                       17
6                        NA                       15
                                                               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 column that details the total number of people fully vaccinated is “persona_fully_vaccinated”.

Q2. What column details the Zip code tabulation area? The column that details the Zip code tabulation area is “zip_code_tabulation_area”.

```
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"
```

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

```
[1] "2022-11-22" "2022-11-22" "2022-11-22" "2022-11-22" "2022-11-22"
[6] "2022-11-22"
```

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 2022-11-22

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

Table 1: Data summary

Name	vax
Number of rows	174636
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	99	0
local_health_jurisdiction	0	1	0	15	495	62	0
county	0	1	0	15	495	59	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
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	8613	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	
age12_plus_population	0	1.00	18895.04	18993.88	0	1346.95	13685.10	31756.12	88556.7	
age5_plus_population	0	1.00	20875.24	21105.98	0	1460.50	15364.00	34877.00	101902.0	
tot_population	8514	0.95	23372.77	22628.51	12	2126.00	18714.00	38168.00	111165.0	
persons_fully_vaccinated	14921	0.91	13466.34	14722.46	11	883.00	8024.00	22529.00	87186.0	
persons_partially_vaccinated	14921	0.91	1707.50	1998.80	11	167.00	1194.00	2547.00	39204.0	
percent_of_population_fully_vaccinated	18665	0.89	0.55	0.25	0	0.39	0.59	0.73	1.0	
percent_of_population_partially_vaccinated	18665	0.89	0.08	0.09	0	0.05	0.06	0.08	1.0	
percent_of_population_with_1_plus_dose	19562	0.89	0.61	0.25	0	0.46	0.65	0.79	1.0	
booster_recip_count	70421	0.60	5655.17	6867.49	11	280.00	2575.00	9421.00	58304.0	
bivalent_dose_recip_count	156958	0.10	1646.02	2161.84	11	109.00	719.00	2443.00	18109.0	
eligible_recipient_count	0	1.00	12309.19	14555.83	0	466.00	5810.00	21140.00	86696.0	

> Q5. How many numeric columns are in this dataset? There are 13 numeric columns in this dataset

```
sum(is.na(vax$persons_fully_vaccinated))
```

```
[1] 14921
```

> Q6. Note that there are "missing values" in the dataset. How many NA values there in the persons_fully_vaccinated column? According to my data set there is 14921 missing values, however in the lab sheet there is 15440 missing.

> Q7. What percent of persons_fully_vaccinated values are missing (to 2 significant figures)? In my data set there is 8.54% percent of persona_fully_vaccinated values missing. In the lab sheet there is 8.93%

```
nrow(vax)
```

```
[1] 174636
```

```r
14921/174636 * 100
```

```
[1] 8.544057
```

> Q8. [Optional]: Why might this data be missing? Data might be missing because the person might of requested to opt out of their information being studied.

Working with Dates

Using the `lubridate` package, dates and times become easier to work with.

```r
library(lubridate)
```

```
Loading required package: timechange


Attaching package: 'lubridate'


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

    date, intersect, setdiff, union
```

```r
today()
```

```
[1] "2022-11-25"
```

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

Now we can do math with dates. For example: How many days have passed since the first vaccination reported in this dataset?

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

```
Time difference of 689 days
```

Using the last and the first date value we can now determine how many days the dataset span.

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

```
Time difference of 686 days
```

> Q9. How many days have passed since the last update of the dataset? According to my dataset only 3 days have passed, whilst the lab sheet 6 days have passed.
>
> Q10. How many unique dates are in the dataset (i.e. how many different dates are detailed)? 99 unique dates according to my dataset, whilst the lab sheet is 97 unique dates.

```
(unique(vax$as_of_date))
```

```
 [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" "2022-03-08" "2022-03-15" "2022-03-22" "2022-03-29"
[66] "2022-04-05" "2022-04-12" "2022-04-19" "2022-04-26" "2022-05-03"
[71] "2022-05-10" "2022-05-17" "2022-05-24" "2022-05-31" "2022-06-07"
[76] "2022-06-14" "2022-06-21" "2022-06-28" "2022-07-05" "2022-07-12"
[81] "2022-07-19" "2022-07-26" "2022-08-02" "2022-08-09" "2022-08-16"
[86] "2022-08-23" "2022-08-30" "2022-09-06" "2022-09-13" "2022-09-20"
[91] "2022-09-27" "2022-10-04" "2022-10-11" "2022-10-18" "2022-10-25"
[96] "2022-11-01" "2022-11-08" "2022-11-15" "2022-11-22"
```

Working with ZIP codes

In R we can use the zipcodeR package to make working with these codes easier. For example, let's install and then load up this package and to find the centroid of the La Jolla 92037 (i.e. UC San Diego) ZIP code area.

```
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, e.g.

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

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

More usefully, we can pull census data about ZIP code areas (including median household income etc.). For example:

```
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, ...
```

Let's now focus in on the San Diego County area by restricting ourselves first to `vax$county == "San Diego"` entries

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

Using dplyr the code would look like this:

```r
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
```

```r
sd <- filter(vax, county == "San Diego")

nrow(sd)
```

```
[1] 10593
```

Using dplyr is often more convenient when we are subsetting across multiple criteria - for example all San Diego county areas with a population of over 10,000.

```r
sd.10 <- filter(vax, county == "San Diego" &
                age5_plus_population > 10000)
```

> Q11. How many distinct zip codes are listed for San Diego County 107 distinct zip codes are listed for San Diego County.

```r
length(unique(sd$zip_code_tabulation_area))
```

```
[1] 107
```

> Q12. What San Diego County Zip code area has the largest 12 + Population in this dataset? 92154 has the highest 12 + population in this dataset.

```r
which.max(sd.10$age12_plus_population)
```

```
[1] 32
```

```r
sd.10[32,]
```

```
   as_of_date zip_code_tabulation_area local_health_jurisdiction    county
32 2021-01-05                    92154                 San Diego San Diego
   vaccine_equity_metric_quartile                 vem_source
32                              2 Healthy Places Index Score
   age12_plus_population age5_plus_population tot_population
32               76365.2                82971          88979
   persons_fully_vaccinated persons_partially_vaccinated
32                       17                         1379
   percent_of_population_fully_vaccinated
32                               0.000191
   percent_of_population_partially_vaccinated
32                                   0.015498
   percent_of_population_with_1_plus_dose booster_recip_count
32                               0.015689                  NA
   bivalent_dose_recip_count eligible_recipient_count
32                        NA                       17
                                                                redacted
32 Information redacted in accordance with CA state privacy requirements
```

```r
sd.toplot <- filter(vax, county == "San Diego" &
                      as_of_date == "2022-11-15")
```

> Q13. What is the overall average “Percent of Population Fully Vaccinated” value for all San Diego “County” as of “2022-11-15”? The overal average of percent in my dataset is 0.7369099 but in the lab it is “0.738176464646465

```r
mean(na.omit(sd.toplot$percent_of_population_fully_vaccinated))
```

```
[1] 0.7369099
```

> 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”?

```r
library(ggplot2)

##ggplot(sd.toplot, aes(+
     #    geom_bar(stat = "bin")
```