class 17: Vaccination mini project

Jennifer

Let’s start by downlaoding our data and reading/importing it into the object “vax”.

```
# 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                    93562            San Bernardino  San Bernardino
2 2021-01-05                    93437             Santa Barbara   Santa Barbara
3 2021-01-05                    93445           San Luis Obispo San Luis Obispo
4 2021-01-05                    93442           San Luis Obispo San Luis Obispo
5 2021-01-05                    93444           San Luis Obispo San Luis Obispo
6 2021-01-05                    93453           San Luis Obispo San Luis Obispo
  vaccine_equity_metric_quartile                 vem_source
1                              1 Healthy Places Index Score
2                             NA            No VEM Assigned
3                              2 Healthy Places Index Score
4                              3 Healthy Places Index Score
5                              3 Healthy Places Index Score
6                              3 Healthy Places Index Score
  age12_plus_population age5_plus_population tot_population
1                1469.5                 1668           1771
2                2494.5                 2871           3387
3                6116.7                 6762           7106
4               10005.2                10615          10917
5               18951.8                20522          21331
6                2373.6                 2499           2578
  persons_fully_vaccinated persons_partially_vaccinated
1                       NA                           NA
2                       NA                           NA
3                       NA                           NA
```

```
4                       NA                           NA
5                       NA                           NA
6                       NA                           NA
  percent_of_population_fully_vaccinated
1                                     NA
2                                     NA
3                                     NA
4                                     NA
5                                     NA
6                                     NA
  percent_of_population_partially_vaccinated
1                                         NA
2                                         NA
3                                         NA
4                                         NA
5                                         NA
6                                         NA
  percent_of_population_with_1_plus_dose booster_recip_count
1                                     NA                  NA
2                                     NA                  NA
3                                     NA                  NA
4                                     NA                  NA
5                                     NA                  NA
6                                     NA                  NA
  bivalent_dose_recip_count eligible_recipient_count
1                        NA                        0
2                        NA                        1
3                        NA                        0
4                        NA                        1
5                        NA                        0
6                        NA                        0
                                                               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

Q3. What is the earliest date in this dataset?

2021-01-05

Q4. What is the latest date in this dataset?

```r
tail(vax, n=1)
```

```
       as_of_date zip_code_tabulation_area local_health_jurisdiction county
172872 2022-11-15                    95746                    Placer Placer
       vaccine_equity_metric_quartile                 vem_source
172872                              4 Healthy Places Index Score
       age12_plus_population age5_plus_population tot_population
172872               20588.8                22923          23934
       persons_fully_vaccinated persons_partially_vaccinated
172872                    16979                         1108
       percent_of_population_fully_vaccinated
172872                               0.709409
       percent_of_population_partially_vaccinated
172872                                   0.046294
       percent_of_population_with_1_plus_dose booster_recip_count
172872                               0.755703               11492
       bivalent_dose_recip_count eligible_recipient_count redacted
172872                      3809                    16877       No
```

2022-11-15

Let's call the skim() function from the skimr package to get a quick overview of this dataset:

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

Table 1: Data summary

Name	vax
Number of rows	172872
Number of columns	18
Column type frequency:	
character	5
numeric	13

Table 1: Data summary

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	98	0
local_health_jurisdiction	0	1	0	15	490	62	0
county	0	1	0	15	490	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	8526	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	8428	0.95	23372.77	22628.51	12	2126.00	18714.00	38168.00	111165.0	
persons_fully_vaccinated	15440	0.91	13309.15	14740.07	11	859.00	7687.00	22253.00	87305.0	
persons_partially_vaccinated	15440	0.91	1679.13	1993.86	11	157.00	1158.00	2483.00	39201.0	
percent_of_population_fully_vaccinated	18986	0.89	0.54	0.26	0	0.36	0.58	0.73	1.0	
percent_of_population_partially_vaccinated	18986	0.89	0.08	0.09	0	0.05	0.06	0.08	1.0	
percent_of_population_with_1_plus_dose	19822	0.89	0.60	0.26	0	0.42	0.64	0.79	1.0	
booster_recip_count	70642	0.59	5701.06	6972.68	11	276.00	2546.00	9513.00	58301.0	
bivalent_dose_recip_count	156937	0.09	1512.94	1994.71	11	101.00	662.00	2236.00	16790.0	
eligible_recipient_count	0	1.00	12114.81	14551.97	0	438.00	5520.00	20714.00	86817.0	

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

13

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

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

```
[1] 15440
```

15400

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

8.93

The “lubridate” package will help us dates and times.

```
library(lubridate)
```

```
Loading required package: timechange
```

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

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

    date, intersect, setdiff, union
```

```
today()
```

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

We can do math with dates by converting our date data into a lubridate format.

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

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

```
Time difference of 686 days
```

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

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

```
Time difference of 679 days
```

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

6

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

98

In R we can use the zipcodeR package to make working with these codes easier

```r
library(zipcodeR)
```

We can use the dplyr package to focus in on the San Diego County area.

```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] 10486
```

```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

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

```r
which.max("age12_plus_population")
```

```
Warning in which.max("age12_plus_population"): NAs introduced by coercion
```

```
integer(0)
```

92154

UC San Diego resides in the 92037 ZIP code area and is listed with an age 5+ population size of 36,144.

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

```
[1] 36144
```