

Class 17 Lab

Getting Started

```
file <- "covid.csv"
vax <- read.csv(file)
head(vax)
```

```
as_of_date zip_code_tabulation_area local_health_jurisdiction county
1 2021-01-05 91606 Los Angeles Los Angeles
2 2021-01-05 95312 Merced Merced
3 2021-01-05 91350 Los Angeles Los Angeles
4 2021-01-05 91708 San Bernardino San Bernardino
5 2021-01-05 95305 Tuolumne Tuolumne
6 2021-01-05 91351 Los Angeles Los Angeles
vaccine_equity_metric_quartile vem_source
1 1 Healthy Places Index Score
2 1 CDPH-Derived ZCTA Score
3 4 Healthy Places Index Score
4 NA No VEM Assigned
5 NA No VEM Assigned
6 3 Healthy Places Index Score
age12_plus_population age5_plus_population tot_population
1 38210.0 41964 44295
2 187.4 236 276
3 29940.2 33775 36173
4 3517.3 3794 NA
5 0.0 0 NA
6 27874.9 30641 32711
persons_fully_vaccinated persons_partially_vaccinated
1 14 482
2 NA NA
```

3	65	1225
4	NA	NA
5	NA	NA
6	31	644
percent_of_population_fully_vaccinated		
1	0.000316	
2	NA	
3	0.001797	
4	NA	
5	NA	
6	0.000948	
percent_of_population_partially_vaccinated		
1	0.010882	
2	NA	
3	0.033865	
4	NA	
5	NA	
6	0.019688	
percent_of_population_with_1_plus_dose booster_recip_count		
1	0.011198	NA
2	NA	NA
3	0.035662	NA
4	NA	NA
5	NA	NA
6	0.020636	NA
bivalent_dose_recip_count eligible_recipient_count		
1	NA	14
2	NA	0
3	NA	65
4	NA	6
5	NA	0
6	NA	31
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?

2022-11-29

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

Table 1: Data summary

Name	vax
Number of rows	176400
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	100	0
local_health_jurisdiction	0	1	0	15	500	62	0
county	0	1	0	15	500	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	0	192257.75	3658.50	5380.50	7635.0	
vaccine_equity_metric_0700tile	0	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	
age12_plus_population	0	1.00	18895.04	8993.88	0	1346.95	13685.13	1756.18	8556.7	

skim_variable	n_missing	complete	mean	sd	p0	p25	p50	p75	p100	hist
age5_plus_population	0	1.00	20875.22	1105.98	0	1460.50	5364.00	14877.00	101902.0	
tot_population	8600	0.95	23372.77	2628.51	2	2126.00	18714.00	18168.00	11165.0	
persons_fully_vaccinated	15048	0.91	13504.90	4748.88	1	887.00	8076.00	22588.00	87207.0	
persons_partially_vaccinated	15048	0.91	1707.77	2001.11	11	167.00	1195.00	2547.00	39228.0	
percent_of_population_fully_vaccinated	18834	0.89	0.55	0.25	0	0.40	0.59	0.73	1.0	
percent_of_population_partially_vaccinated	18834	0.89	0.08	0.09	0	0.05	0.06	0.08	1.0	
percent_of_population_1_plus_dose	19739	0.89	0.62	0.25	0	0.46	0.65	0.79	1.0	
booster_recip_count	70611	0.60	5643.35	6858.00	11	281.00	2585.00	9377.00	58376.0	
bivalent_dose_recip_count	157094	0.11	1770.66	2315.50	11	117.00	778.00	2643.75	18815.0	
eligible_recipient_count	0	1.00	12345.64	4582.42	0	468.00	5851.00	21198.25	56706.0	

Q5. How many numeric columns are in this dataset

13 numeric columns

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

15048 NA values

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

0.09

Working with Dates

```
library(lubridate)
```

Loading required package: timechange

Attaching package: 'lubridate'

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

date, intersect, setdiff, union

```
today()
```

```
[1] "2022-12-05"
```

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

Time difference of 699 days

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

Time difference of 693 days

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

6 days

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

```
date <- vax$as_of_date
date_unique <- unique(date)
length(date_unique)
```

```
[1] 100
```

100 unique days

Working with Zip Codes

```
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_~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, ...
```

Focus on the San Diego Area

```
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] 10700

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

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

```
zip <- sd$zip_code_tabulation_area  
zip_unique <- unique(zip)  
length(zip_unique)
```

[1] 107

There are 107 distinct zip codes in SD county.

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

```
age12 <- sd$age12_plus_population  
which.max(age12)
```

[1] 77

```
sd[77,2]
```

[1] 92154

92154 has the largest 12+ population in the dataset