R Notebook

Saumya Ranjan (A15483401)

11/23/2021

```
# Import vaccination data
vax <- read.csv("covid19vaccinesbyzipcode_test.csv")
head(vax)
```

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 1 2021-01-05                   92804                     Orange    Orange
## 2 2021-01-05                   92626                     Orange    Orange
## 3 2021-01-05                   92250                   Imperial  Imperial
## 4 2021-01-05                   92637                     Orange    Orange
## 5 2021-01-05                   92155                  San Diego San Diego
## 6 2021-01-05                   92259                   Imperial  Imperial
##   vaccine_equity_metric_quartile                vem_source
## 1                              2 Healthy Places Index Score
## 2                              3 Healthy Places Index Score
## 3                              1 Healthy Places Index Score
## 4                              3 Healthy Places Index Score
## 5                             NA            No VEM Assigned
## 6                              1    CDPH-Derived ZCTA Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1               76455.9                84200                       19
## 2               44238.8                47883                       NA
## 3                7098.5                 8026                       NA
## 4               16027.4                16053                       NA
## 5                 456.0                  456                       NA
## 6                 119.0                  121                       NA
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                         1282                                0.000226
## 2                           NA                                      NA
## 3                           NA                                      NA
## 4                           NA                                      NA
## 5                           NA                                      NA
## 6                           NA                                      NA
##   percent_of_population_partially_vaccinated
## 1                                   0.015226
## 2                                         NA
## 3                                         NA
## 4                                         NA
## 5                                         NA
## 6                                         NA
##   percent_of_population_with_1_plus_dose
## 1                               0.015452
## 2                                     NA
## 3                                     NA
```

```
## 4                                       NA
## 5                                       NA
## 6                                       NA
##                                                                    redacted
## 1                                                                        No
## 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

##Ensure Date column is useful

```
library(lubridate)
```

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

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

```
today()
```

```
## [1] "2021-11-23"
```

Here we make our 'as_of_date' column lubridate format

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

Now I can do math with dates more easily

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

```
## Time difference of 322 days
```

How many days since last entry?

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

```
## Time difference of 7 days
```

Q9. How many days between the first and last entry in the dataset

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

```
## Time difference of 315 days
```

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

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

```
## [1] 46
```

This makes sense because

```
46 * 7
```

```
## [1] 322
```

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

Table 1: Data summary

Name	vax
Number of rows	81144
Number of columns	14
Column type frequency:	
character	4
Date	1
numeric	9
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
local_health_jurisdiction	0	1	0	15	230	62	0
county	0	1	0	15	230	59	0
vem_source	0	1	15	26	0	3	0
redacted	0	1	2	69	0	2	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
as_of_date	0	1	2021-01-05	2021-11-16	2021-06-11	46

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	1002	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	
age12_plus_population	0	1.00	18895.04	18993.94	0	1346.95	13685.10	31756.12	88556.7	
age5_plus_population	0	1.00	20875.24	21106.05	0	1460.50	15364.00	34877.00	101902.0	
persons_fully_vaccinated	8256	0.90	9456.49	11498.25	11	506.00	4105.00	15859.00	71078.0	
persons_partially_vaccinated	8256	0.90	1900.61	2113.07	11	200.00	1271.00	2893.00	20185.0	
percent_of_population_fully_vaccinated	8256	0.90	0.42	0.27	0	0.19	0.44	0.62	1.0	
percent_of_population_partially_vaccinated	8256	0.90	0.10	0.10	0	0.06	0.07	0.11	1.0	
percent_of_population_with_12plus_dose	8256	0.90	0.50	0.26	0	0.30	0.53	0.70	1.0	

Q5. How many numeric columns are in this dataset?

–14

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

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

```
## [1] 8256
```

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

```
 num <- sum (is.na (vax$persons_fully_vaccinated)/ nrow(vax)*100)

round(num, 2)
```

```
## [1] 10.17
```

```
#library(zipcodeR)
```

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

Working with ZIP codes

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

#Focus on the San Diego area

```
inds <- vax$county == "San Diego"
```

```
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] 4922
```

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

Q11.

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

```
## [1] 107
```

Q12

```
ind <- which.max(sd$age12_plus_population)
sd[ind,]
```

```
##    as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 23 2021-01-05                   92154                San Diego San Diego
```

```
##    vaccine_equity_metric_quartile                 vem_source
## 23                              2 Healthy Places Index Score
##    age12_plus_population age5_plus_population persons_fully_vaccinated
## 23               76365.2                82971                       32
##    persons_partially_vaccinated percent_of_population_fully_vaccinated
## 23                         1336                                0.000386
##    percent_of_population_partially_vaccinated
## 23                                   0.016102
##    percent_of_population_with_1_plus_dose redacted
## 23                                0.016488       No
```

92154

Q13

```
sd.now <- filter(sd, as_of_date == "2021-11-09")

mean(sd.now$percent_of_population_fully_vaccinated, na.rm=TRUE)
```

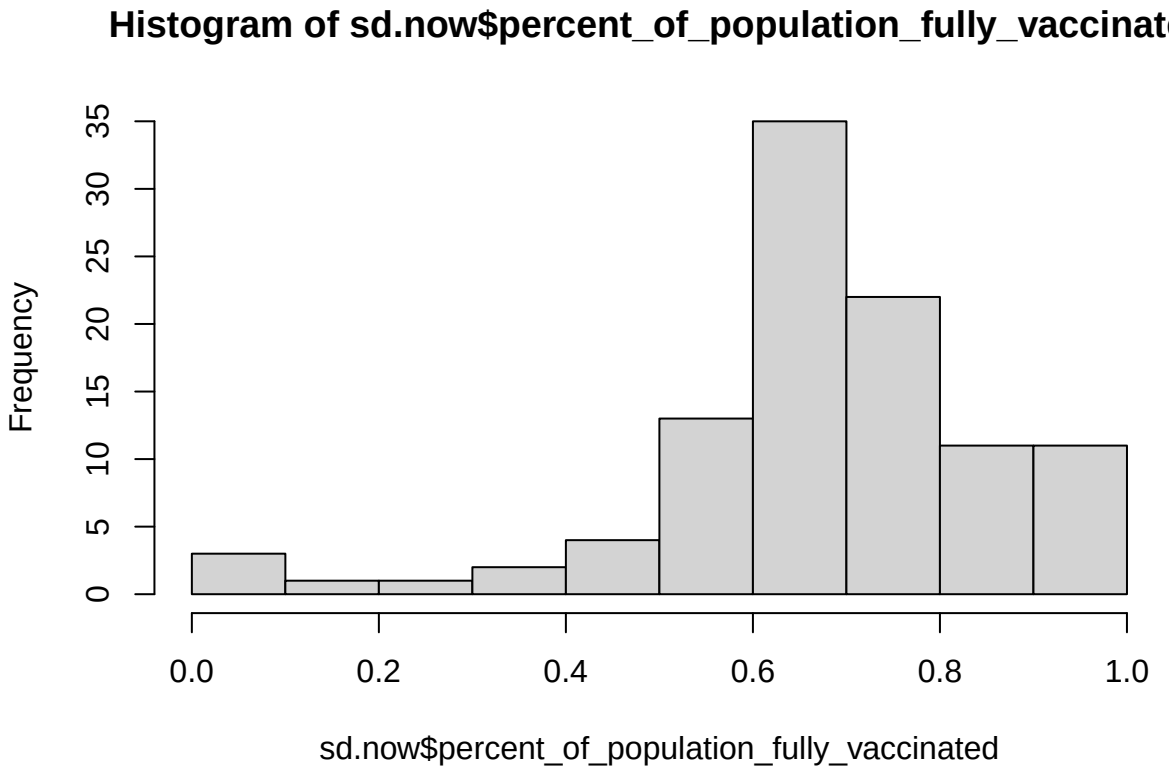
```
## [1] 0.6727567
```

We can look at the 6 number summary

```
summary(sd.now$percent_of_population_fully_vaccinated)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.01017 0.60776 0.67700 0.67276 0.76164 1.00000       4
```

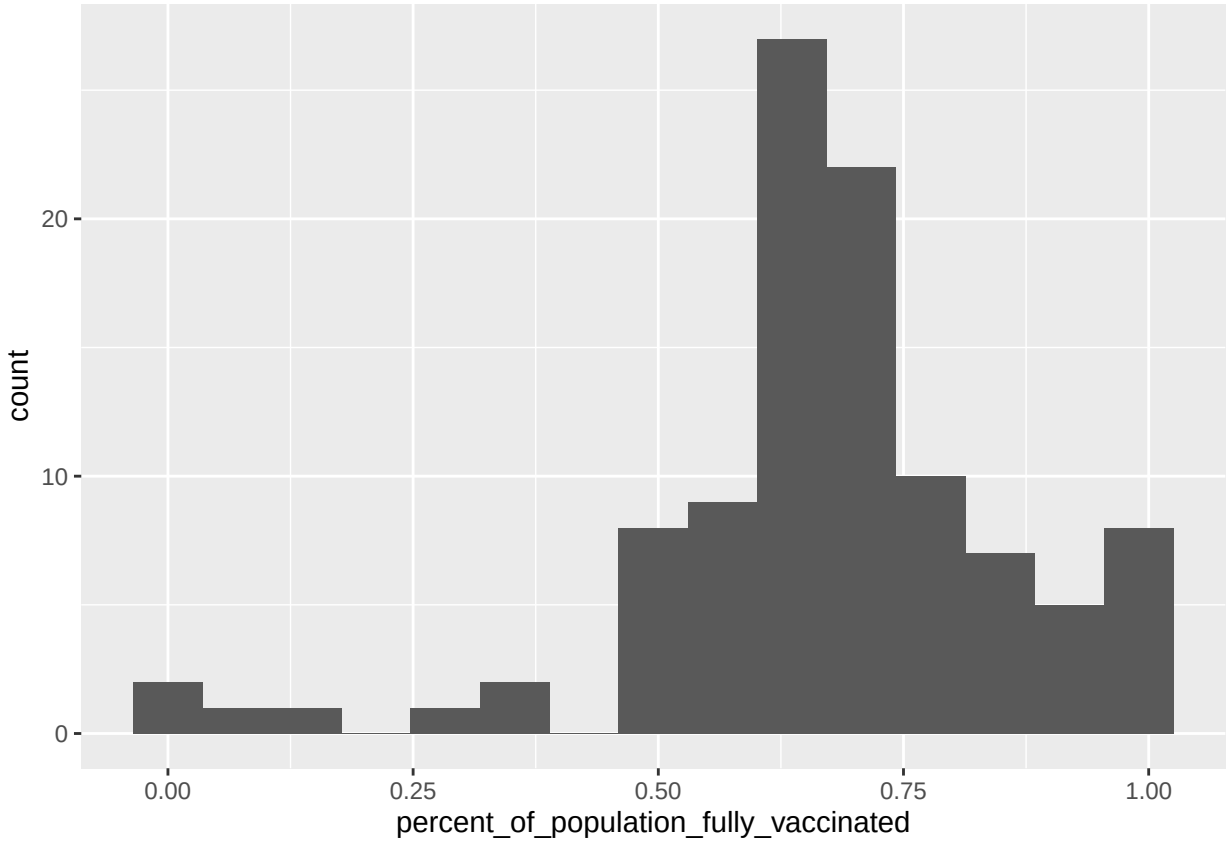
```
hist(sd.now$percent_of_population_fully_vaccinated)
```



```
library(ggplot2)

ggplot(sd.now) + aes(percent_of_population_fully_vaccinated) + geom_histogram(bins=15)
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```




```
#What about 90237 La Jolla/ UCSD

ucsd <- filter(sd, zip_code_tabulation_area =="92037")
head(ucsd)
```

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 1 2021-01-05                   92037                San Diego San Diego
## 2 2021-01-12                   92037                San Diego San Diego
## 3 2021-01-19                   92037                San Diego San Diego
## 4 2021-01-26                   92037                San Diego San Diego
## 5 2021-02-02                   92037                San Diego San Diego
## 6 2021-02-09                   92037                San Diego San Diego
##   vaccine_equity_metric_quartile               vem_source
## 1                              4 Healthy Places Index Score
## 2                              4 Healthy Places Index Score
## 3                              4 Healthy Places Index Score
## 4                              4 Healthy Places Index Score
## 5                              4 Healthy Places Index Score
## 6                              4 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1               33675.6                36144                       44
```

```
## 2               33675.6               36144                      470
## 3               33675.6               36144                      730
## 4               33675.6               36144                     1079
## 5               33675.6               36144                     1616
## 6               33675.6               36144                     2222
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                         1265                                0.001217
## 2                         1565                                0.013004
## 3                         3505                                0.020197
## 4                         6197                                0.029853
## 5                         8388                                0.044710
## 6                         9634                                0.061476
##   percent_of_population_partially_vaccinated
## 1                                   0.034999
## 2                                   0.043299
## 3                                   0.096973
## 4                                   0.171453
## 5                                   0.232072
## 6                                   0.266545
##   percent_of_population_with_1_plus_dose redacted
## 1                               0.036216       No
## 2                               0.056303       No
## 3                               0.117170       No
## 4                               0.201306       No
## 5                               0.276782       No
## 6                               0.328021       No
```

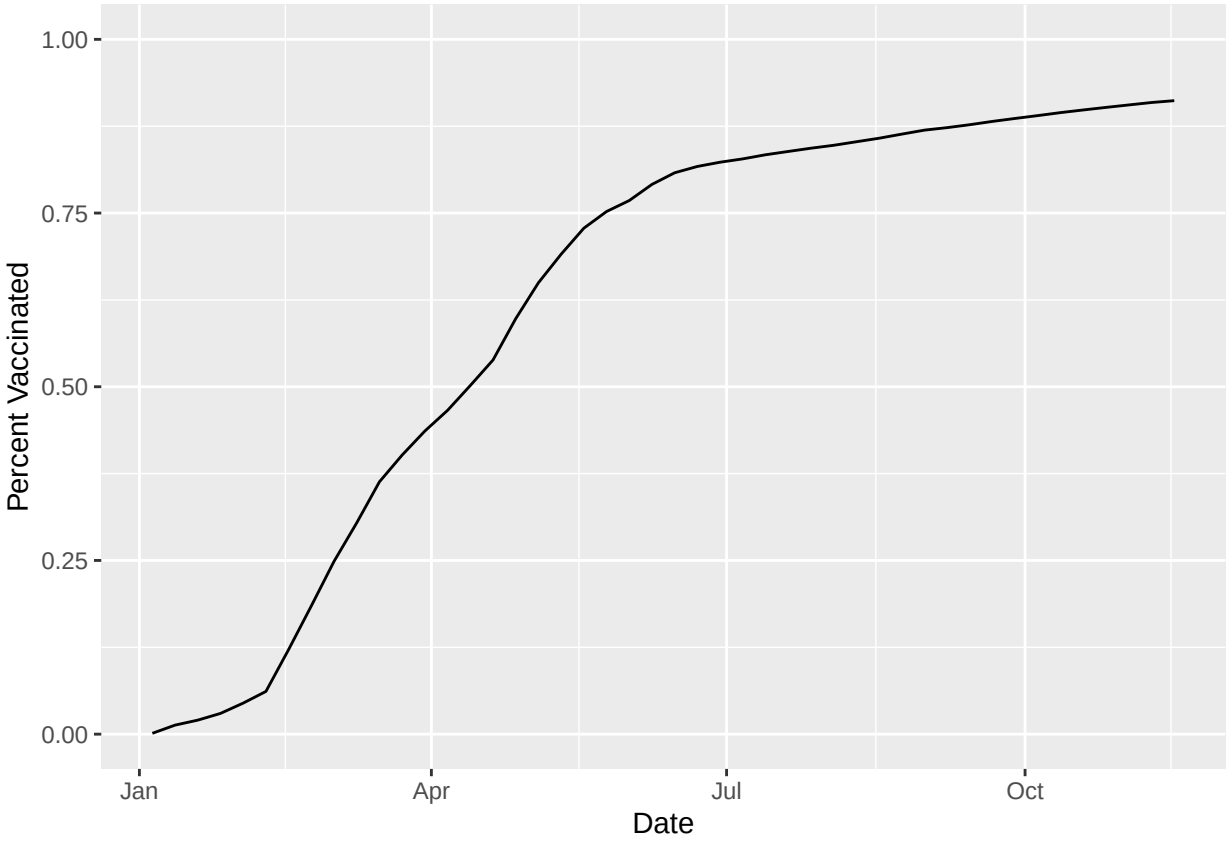
```
ucsd[1,]$age5_plus_population
```

```
## [1] 36144
```

Time series of vaccination rate for 92037

```
ggplot(ucsd) +
  aes(as_of_date,
      percent_of_population_fully_vaccinated) +
  geom_line(group=1) +
  ylim(c(0,1)) +
  labs(x="Date", y="Percent Vaccinated")+
  geom_hline(yintercept =66 , col="red")
```

```
## Warning: Removed 1 rows containing missing values (geom_hline).
```



Population in the 92037 ZIP Code area

```
ucsd[1,]$age5_plus_population
```

```
## [1] 36144
```

First we need to subset the full 'vax' dataset to include only zipcode areas with a population as large as 92037

```
vax.36.all <- filter(vax, age5_plus_population > 36144)
nrow(vax.36.all)
```

```
## [1] 18906
```

How many unique zipcodes have a population as large as 92037

```
length(unique (vax.36.all$zip_code_tabulation_area))
```

```
## [1] 411
```

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 == "2021-11-16") %>%
  filter(zip_code_tabulation_area=="92040") %>%
  select(percent_of_population_fully_vaccinated)
```

```
##   percent_of_population_fully_vaccinated
## 1                               0.520463
```

below

Lets show a final figure that shows all of these zipcodes

```
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") +
   labs(x="Date", y="Percent Vaccinated",
        title="Vaccination Rate Across California",
        subtitle="Only areas with population above 36k are shown") +
  geom_hline(yintercept = 0.66, col="red")
```

```
## Warning: Removed 180 row(s) containing missing values (geom_path).
```