Class 17 Report

Divyanshu Kawankar

11/29/2021

###Getting Started

```
vax <- read.csv("15702a90-aa5d-49bc-8621-a8129630725a.csv")

head(vax)
```

	as_of_date <chr>	zip_code_tabulation_area <int>	local_health_jurisdiction <chr>	county <chr>
1	2021-01-05	92395	San Bernardino	San Bernardino
2	2021-01-05	93206	Kern	Kern
3	2021-01-05	91006	Los Angeles	Los Angeles
4	2021-01-05	91901	San Diego	San Diego
5	2021-01-05	92230	Riverside	Riverside
6	2021-01-05	92662	Orange	Orange

6 rows | 1-5 of 15 columns

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

```
## [1] "2021-11-23" "2021-11-23" "2021-11-23" "2021-11-23" "2021-11-23"
## [6] "2021-11-23"
```

> 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? - 2021-11-23

```
#Let's get am overview of this dataset
library(skimr)

skimr::skim(vax)
```

Data summary

Name	vax
Number of rows	82908
Number of columns	14
_______________________	
Column type frequency:	
character	5
numeric	9
________________________	
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	47	0
local_health_jurisdiction	0	1	0	15	235	62	0
county	0	1	0	15	235	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	4089	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	21106.04	0	1460.50	15364.00	34877.00	101902.0	▇▃▂
persons_fully_vaccinated	8355	0.90	9585.35	11609.12	11	516.00	4210.00	16095.00	71219.0	▇▂▁
persons_partially_vaccinated	8355	0.90	1894.87	2105.55	11	198.00	1269.00	2880.00	20159.0	▇▁▁
percent_of_population_fully_vaccinated	8355	0.90	0.43	0.27	0	0.20	0.44	0.63	1.0	▇▆▇
percent_of_population_partially_vaccinated	8355	0.90	0.10	0.10	0	0.06	0.07	0.11	1.0	▇▁▁
percent_of_population_with_1_plus_dose	8355	0.90	0.51	0.26	0	0.31	0.53	0.71	1.0	▅▅▇

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

```
## [1] 8355
```

```
sum( is.na(vax$persons_fully_vaccinated) ) / ( sum( is.na(vax$persons_fully_vaccinated) ) + sum( is.na(vax$persons_fully_vac
cinated) == FALSE ) )
```

```
## [1] 0.1007744
```

> Q5. How many numeric columns are in this dataset? - 9

> Q6. Note that there are “missing values” in the dataset. How many NA values there in the persons_fully_vaccinated column? - 8355

> Q7. What percent of persons_fully_vaccinated values are missing (to 2 significant figures)? - 10%

```
library(lubridate)
```

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

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

```
today()
```

```
## [1] "2021-11-29"
```

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

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

```
## Time difference of 328 days
```

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

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

```
today() - vax$as_of_date[82908]
```

```
## Time difference of 6 days
```

```
(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"
```

> 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)? -47 for me but 46 when the prof made it.

###Working with Zip Codes

```
library(zipcodeR)

geocode_zip('92037')
```

zipcode <chr>	lat <dbl>	lng <dbl>
92037	32.8	-117.2

1 row

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

zipcode_a <chr>	zipcode_b <chr>	distance <dbl>
92037	92109	2.33

1 row

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

zipcode <chr>	zipcode_type <chr>	major_city <chr>	post_office_city <chr>	common_city_list <blob>	county <chr>	state <chr>	lat <dbl> ▸
92037	Standard	La Jolla	La Jolla, CA	<blob>	San Diego County	CA	32.8
92109	Standard	San Diego	San Diego, CA	<blob>	San Diego County	CA	32.8

2 rows | 1-8 of 24 columns

```
zipdata <- reverse_zipcode( vax$zip_code_tabulation_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] 5029
```

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

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

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

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

```
## [1] 60
```

```
(sd$zip_code_tabulation[60])
```

```
## [1] 92154
```

> 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? -92154. I did it kind of differently. First found which place was max. Then found the zipcode for that place.

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

```
## [1] 0.4460157
```

> # Q13. What is the overall average “Percent of Population Fully Vaccinated” value for all San Diego “County” as of “2021-11-09”?

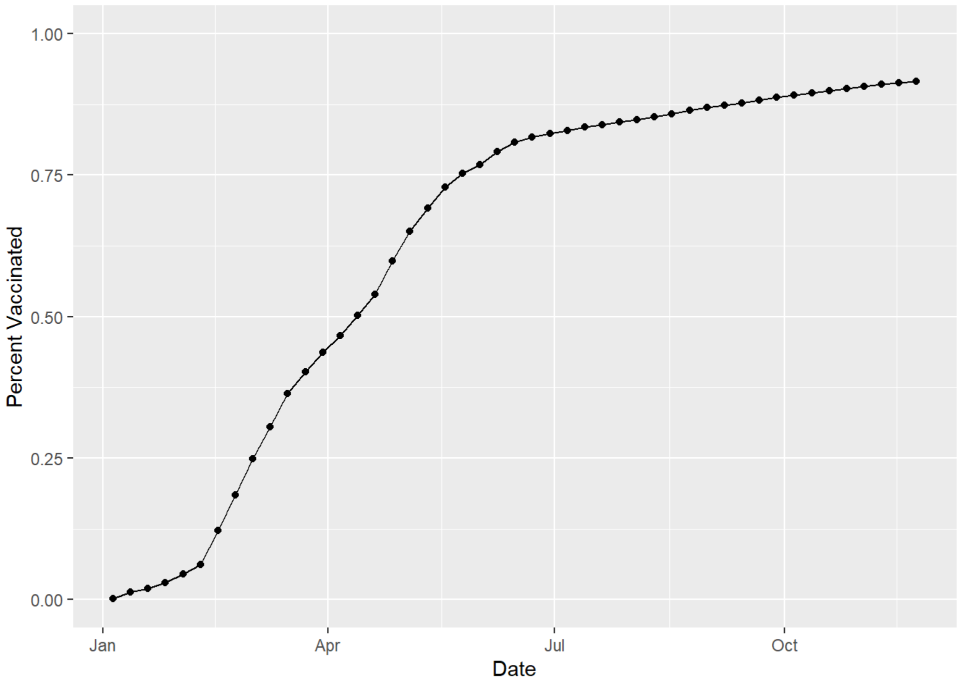
> # Q14. Using either ggplot or base R graphics make a summary figure that shows the distribution of Percent of Population Fully Vaccinated values as of “2021-11-09”?

```
library(ggplot2)
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: - Below

```
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")
```



```
vax.36 <- filter(vax, age5_plus_population > 36144 &
                as_of_date == "2021-11-16")

head(vax.36)
```

	as_of_date	zip_code_tabulation_area	local_health_jurisdiction	county	▶
	<date>	<int>	<chr>	<chr>	
1	2021-11-16	92020	San Diego	San Diego	
2	2021-11-16	92563	Riverside	Riverside	
3	2021-11-16	92806	Orange	Orange	
4	2021-11-16	93291	Tulare	Tulare	
5	2021-11-16	92335	San Bernardino	San Bernardino	

	as_of_date <date>	zip_code_tabulation_area <int>	local_health_jurisdiction <chr>	county <chr>	▸
6	2021-11-16	92618	Orange	Orange	

6 rows | 1-5 of 15 columns

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

```
## [1] 0.6640413
```

```
mean.pop.vax <- mean(vax.36$percent_of_population_fully_vaccinated)
```

> 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 “2021-11-16”. Add this as a straight horizontal line to your plot from above with the geom_hline() function? - 0.6640413. It’s prob diff from prof due to updated dataset.

> 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 “2021-11-16”? - Refer to table below

```
skimr::skim(mean.pop.vax)
```

Data summary

Name	mean.pop.vax
Number of rows	1
Number of columns	1
________________________	
Column type frequency:	
numeric	1
________________________	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
data	0	1	0.66	NA	0.66	0.66	0.66	0.66	0.66	▁▁▇▁▁

```
vax %>% filter(as_of_date == "2021-11-16") %>%
  filter(zip_code_tabulation_area=="92109") %>%
  select(percent_of_population_fully_vaccinated)
```

percent_of_population_fully_vaccinated <dbl>
0.68863

1 row

```
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 <dbl>

percent_of_population_fully_vaccinated <dbl>
0.521047

1 row

> Q19. Is the 92109 and 92040 ZIP code areas above or below the average value you calculated for all these above? - 92109 zip code is higher but the 92040 is lower.