Class17: Mini Project

Katie (PID A15849062)

We will start by downloading the most recently dated "Statewide COVID-19 Vaccines Administered by ZIP Code" CSV file from: https://data.ca.gov/dataset/covid-19-vaccine-progress-dashboard-data-by-zip-code

Import vaccination data

```r
vax <- read.csv("covid19vaccinesbyzipcode_test.csv")
head(vax)
```

```
  as_of_date zip_code_tabulation_area local_health_jurisdiction    county
1 2021-01-05                    95446                    Sonoma    Sonoma
2 2021-01-05                    96014                  Siskiyou  Siskiyou
3 2021-01-05                    96087                    Shasta    Shasta
4 2021-01-05                    96008                    Shasta    Shasta
5 2021-01-05                    95410                 Mendocino Mendocino
6 2021-01-05                    95527                   Trinity   Trinity
  vaccine_equity_metric_quartile                 vem_source
1                              2 Healthy Places Index Score
2                              2    CDPH-Derived ZCTA Score
3                              2    CDPH-Derived ZCTA Score
4                             NA            No VEM Assigned
5                              3    CDPH-Derived ZCTA Score
6                              2    CDPH-Derived ZCTA Score
  age12_plus_population age5_plus_population tot_population
1                4840.7                 5057           5168
2                 135.0                  135            135
3                 513.9                  544            544
4                1125.3                 1164             NA
5                 926.3                  988            997
```

```
6                476.6                  485            499
  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                        0
3                        NA                        2
4                        NA                        2
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
```

```r
head(vax$persons_fully_vaccinated)
```

```
[1] NA NA NA NA NA NA
```

> 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

```r
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"
```

> Q3. What is the earliest date in this dataset?

2021-01-05

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

```
[1] "2023-02-28"
```

> Q4. What is the latest date in this dataset?

2023-02-28

We can use the skim() function for a quick overview.

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

Table 1: Data summary

Name	vax
Number of rows	199332
Number of columns	18
Column type frequency:	
character	5
numeric	1

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	113	0
local_health_jurisdiction	0	1	0	15	565	62	0
county	0	1	0	15	565	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	93665.11	1817.38	90001	92257.75	93658.5	95380.5	97635	

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

13 columns

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

```
[1] 16525
```

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

16525

```r
n.missing <- sum(is.na(vax$persons_fully_vaccinated))
n.missing
```

```
[1] 16525
```

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

```r
round((n.missing / nrow(vax))*100, 2)
```

```
[1] 8.29
```

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

They may be on federal land, and the military does not report this health information.

##Working with dates

The lubridate package makes working with dates and times in R much less of a pain. Let's have a first play with this package here.

```r
library(lubridate)
```

```

Attaching package: 'lubridate'

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

    date, intersect, setdiff, union
```

```r
today()
```

```
[1] "2023-03-12"
```

We can magically do math with dates

```r
today() - ymd("2021-01-05")
```

```
Time difference of 796 days
```

How old am I?

```r
today() - ymd("2001-04-17")
```

```
Time difference of 7999 days
```

Let's treat the whole column as date format

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

Q. How many days have passed since the first vaccination reported in this dataset?

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

```
Time difference of 796 days
```

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

```
Time difference of 784 days
```

Q.9 How many days ago was the data set updated?

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

```
Time difference of 12 days
```

Q.10 How many unique dates are in the dataset (i.e. how many different dates are detailed)?

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

```
[1] 113
```

##Working with Zip Codes

Zip codes are also rather annoying things to work with as they are numeric but not in the conventional sense of doing math.

Just like dates we have special packages to help us work with ZIP codes.

```r
library(zipcodeR)
```

```r
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
```

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

```
head(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

Let's now focus in on the San Diego County area by restricting ourselves first to vax$county == "San Diego" entries. We have two main choices on how to do this. The first using base R the second using the dplyr package:

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

```
[1] 12091
```

It is time to revisit the most awesome **dplyr** package.

```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
library(dplyr)

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

nrow(sd)
```

```
[1] 12091
```

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

```r
n_distinct(sd.10$zip_code_tabulation_area)
```

```
[1] 76
```

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

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

```
[1] 92154
```

```r
head(reverse_zipcode("92154"))
```

```
# A tibble: 1 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 92154   Standard   San Di~ San Di~ <raw 21 B> San D~ CA     32.6  -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, ...
```

> Q13. What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of "2022-11-15"?

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

```
[1] "2023-02-28"
```

> Q13. What is the overall average "Percent of Population Fully Vaccinated" value for all San Diego "County" as of THE MOST RECENT DATE "2023-02-28"

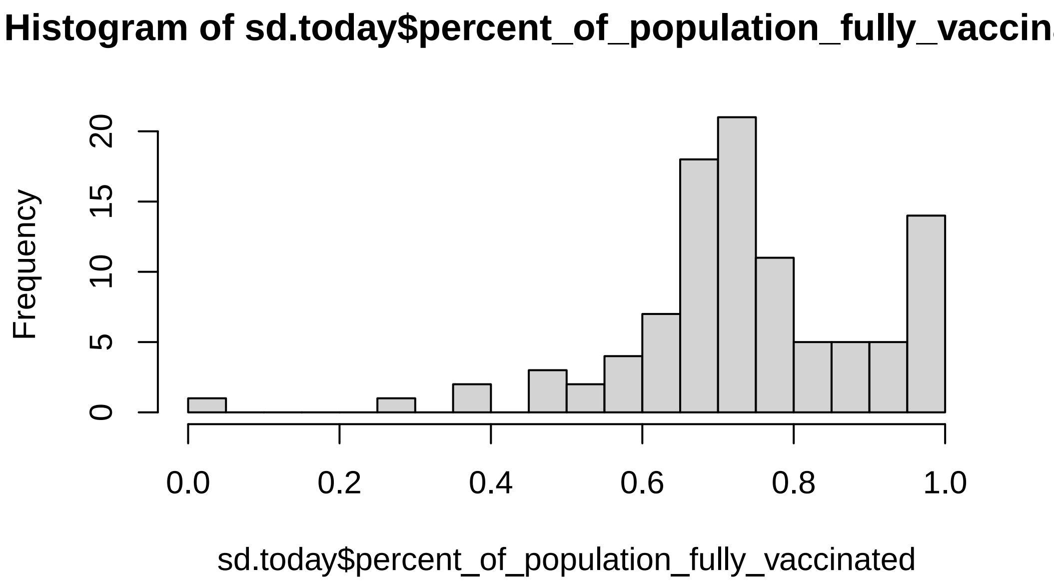
```r
##sd$as_of_date
sd.today <- filter(sd, as_of_date == "2023-02-28")
```

```r
mean(sd.today$percent_of_population_fully_vaccinated, na.rm=T)
```

```
[1] 0.7400878
```

> 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 "2023-02-28"?

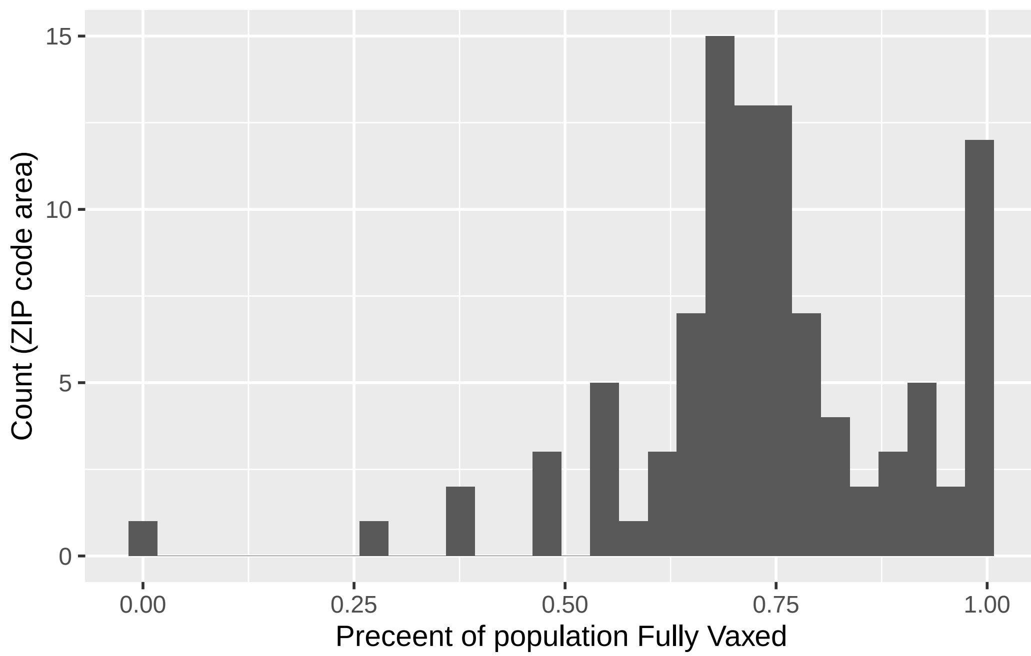
```r
hist(sd.today$percent_of_population_fully_vaccinated, breaks=22)
```




```
library(ggplot2)

ggplot(sd.today)+
  aes(percent_of_population_fully_vaccinated) + geom_histogram() + labs(little="Vacination
  xlab("Preceent of population Fully Vaxed") +
  ylab("Count (ZIP code area)")
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 8 rows containing non-finite values (`stat_bin()`).
```



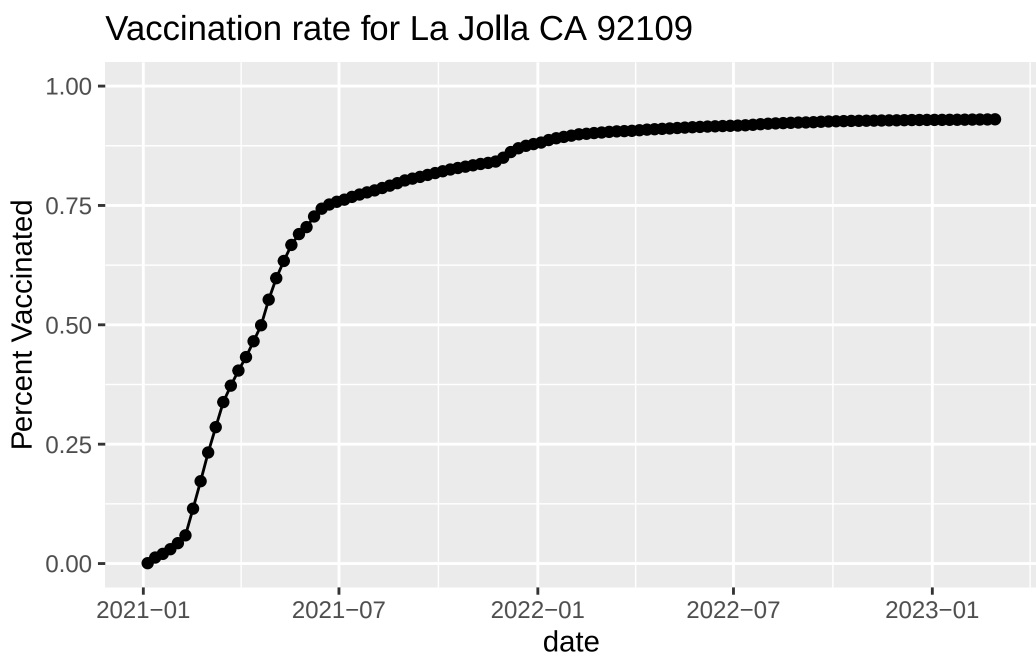
Focus on UCSD?La Jolla

```
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:

```
ucplot <- ggplot(ucsd) +
  aes(as_of_date,
      percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) +
  ylim(c(0,1)) +
  labs(title= "Vaccination rate for La Jolla CA 92109", x="date", y="Percent Vaccinated")
ucplot
```



Comparing to similar sized areas

Let's return to the full dataset and look across every zip code area with a population at least as large as that of 92037 on as_of_date "2022-02-22".

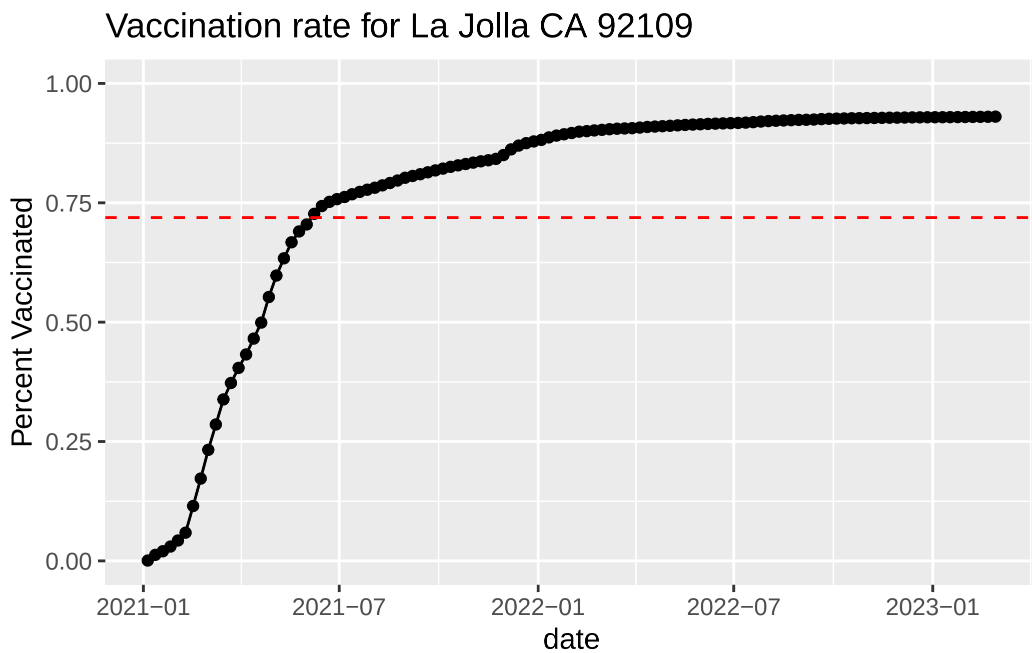
```
# Subset to all CA areas with a population as large as 92037
vax.36 <- filter(vax, age5_plus_population > 36144 &
                as_of_date == "2022-11-15")

#head(vax.36)
```

> 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 "2022-11-15". Add this as a straight horizontal line to your plot from above with the geom_hline() function?

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

```
[1] 0.7190515
```

```r
ucplot + geom_hline(yintercept=ave, col="red", linetype=2)
```



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 "2022-11-15"?

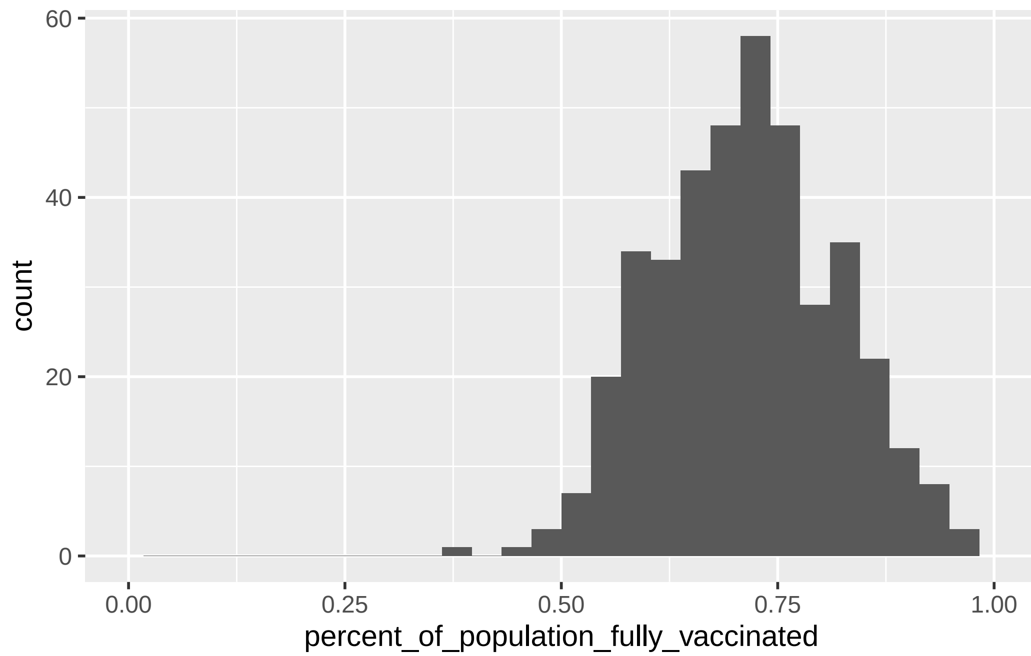
```r
summary(vax.36$percent_of_population_fully_vaccinated)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.3784  0.6444  0.7162  0.7191  0.7882  1.0000 
```

Q18. Using ggplot generate a histogram of this data.

```r
ggplot(vax.36) +
  aes(percent_of_population_fully_vaccinated) +
  geom_histogram() +
  xlim(0,1)
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

Warning: Removed 2 rows containing missing values (`geom_bar()`).
```



Q19. Is the 92109 and 92040 ZIP code areas above or below the average value you calculated for all these above?

```
x <- filter(vax.36, zip_code_tabulation_area %in% c("92109", "92040"))
x$percent_of_population_fully_vaccinated
```

```
[1] 0.548849 0.692874
```

Q20. Finally make a time course plot of vaccination progress for all areas in the full dataset with a age5_plus_population > 36144.

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

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") +
  ylim(0,1) +
  labs(x="Date", y="Percent Vaccinated",
       title="Vaccination rate across California",
       subtitle="Only areas with a population above 36k are shown.") +
```

```
  geom_hline(yintercept = 0.7213, linetype=2)
```

```
Warning: Removed 183 rows containing missing values (`geom_line()`).
```