Lab14

Yushi Li (A15639705)

3/2/2022

Getting started

Data overview

```
# 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                   92549                Riverside      Riverside
## 2 2021-01-05                   92130                San Diego      San Diego
## 3 2021-01-05                   92397           San Bernardino San Bernardino
## 4 2021-01-05                   94563             Contra Costa   Contra Costa
## 5 2021-01-05                   94519             Contra Costa   Contra Costa
## 6 2021-01-05                   91042              Los Angeles    Los Angeles
##   vaccine_equity_metric_quartile               vem_source
## 1                              3 Healthy Places Index Score
## 2                              4 Healthy Places Index Score
## 3                              3 Healthy Places Index Score
## 4                              4 Healthy Places Index Score
## 5                              3 Healthy Places Index Score
## 6                              2 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1                2348.4                 2461                        NA
## 2               46300.3                53102                        61
## 3                3695.6                 4225                        NA
## 4               17216.1                18896                        NA
## 5               16861.2                18678                        NA
## 6               23962.2                25741                        NA
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                           NA                                     NA
## 2                           27                                0.001149
## 3                           NA                                     NA
## 4                           NA                                     NA
## 5                           NA                                     NA
## 6                           NA                                     NA
##   percent_of_population_partially_vaccinated
## 1                                         NA
## 2                                   0.000508
## 3                                         NA
## 4                                         NA
## 5                                         NA
## 6                                         NA
##   percent_of_population_with_1_plus_dose booster_recip_count
## 1                                     NA                  NA
## 2                               0.001657                  NA
## 3                                     NA                  NA
## 4                                     NA                  NA
## 5                                     NA                  NA
## 6                                     NA                  NA
##                                                        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
```

```
# view first and last date
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"
```

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

```
## [1] "2022-03-01" "2022-03-01" "2022-03-01" "2022-03-01" "2022-03-01"
## [6] "2022-03-01"
```

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?

2022-03-01

```
# use skimm
skimr::skim(vax)
```

Data summary

Name	vax
Number of rows	107604
Number of columns	15
Column type frequency:	
character	5
numeric	10
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	61	0
local_health_jurisdiction	0	1	0	15	305	62	0
county	0	1	0	15	305	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	5307	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	▇▁▁▁▇
age12_plus_population	0	1.00	18895.04	18993.91	0	1346.95	13685.10	31756.12	88556.7	▇▃▂▁▂
age5_plus_population	0	1.00	20875.24	21106.02	0	1460.50	15364.00	34877.00	101902.0	▇▃▂▁▂
persons_fully_vaccinated	18338	0.83	12155.61	13063.88	11	1066.25	7374.50	20005.00	77744.0	▇▃▂▁▁

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
persons_partially_vaccinated	18338	0.83	831.74	1348.68	11	76.00	372.00	1076.00	34219.0	
percent_of_population_fully_vaccinated	18338	0.83	0.51	0.26	0	0.33	0.54	0.70	1.0	
percent_of_population_partially_vaccinated	18338	0.83	0.05	0.09	0	0.01	0.03	0.05	1.0	
percent_of_population_with_1_plus_dose	18338	0.83	0.54	0.28	0	0.36	0.58	0.75	1.0	
booster_recip_count	64317	0.40	4100.55	5900.21	11	176.00	1136.00	6154.50	50602.0	

```
# find out how many values r na
sum(is.na(vax$persons_fully_vaccinated))
```

```
## [1] 18338
```

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

```
## [1] 17.04212
```

Q5. How many numeric columns are in this dataset?

10

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

18338

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

17.04%

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

The data might not be collected daily in the 14-month period. It appears to be updated on a weekly basis.

Working with dates

```
# Load the package
library(lubridate)
```

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

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

```
# check today's date
today()
```

```
## [1] "2022-03-03"
```

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

# now we can do math with dates!
today() - vax$as_of_date[1]
```

```
## Time difference of 422 days
```

```
# number of days that the dataset spans:
vax$as_of_date[nrow(vax)] - vax$as_of_date[1]
```

```
## Time difference of 420 days
```

```
# days since last update:
today() - vax$as_of_date[nrow(vax)]
```

```
## Time difference of 2 days
```

```
# number of unique days in dataset:
length(unique(vax$as_of_date))
```

```
## [1] 61
```

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

1 day

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

61 unique dates

Working with ZIP codes

```
# load package
library(zipcodeR)

# test run
geocode_zip('92037')
```

```
## # A tibble: 1 x 3
##   zipcode   lat   lng
##   <chr>   <dbl> <dbl>
## 1 92037    32.8 -117.
```

```
# calculate distance between 2 areas (in miles)
zip_distance('92037','92109')
```

```
##   zipcode_a zipcode_b distance
## 1     92037     92109     2.33
```

```
# get census data from areas
reverse_zipcode(c('92037', "92109"))
```

```
## # A tibble: 2 x 24
##   zipcode zipcode_type major_city post_office_city common_city_list county state
##   <chr>   <chr>        <chr>      <chr>                    <blob> <chr>  <chr>
## 1 92037   Standard     La Jolla   La Jolla, CA          <raw 20 B> San D~ CA
## 2 92109   Standard     San Diego  San Diego, CA         <raw 21 B> San D~ CA
## # ... with 17 more variables: lat <dbl>, lng <dbl>, timezone <chr>,
## #   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>
```

```
# we can pull the data for ALL ZIP codes:
zipdata <- reverse_zipcode( vax$zip_code_tabulation_area )
```

Focus on the San Diego area

San Diego County at large

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

```
## [1] 6527
```

```
# do the same but with dplyr
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] 6527
```

```
#subset over multiple criteria using dplyr
sd.10 <- filter(vax, county == "San Diego" &
                age5_plus_population > 10000)
```

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

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

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

```
head(sd)
```

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 1 2021-01-05                   92130                San Diego San Diego
## 2 2021-01-05                   91945                San Diego San Diego
## 3 2021-01-05                   91917                San Diego San Diego
## 4 2021-01-05                   92103                San Diego San Diego
## 5 2021-01-05                   92075                San Diego San Diego
## 6 2021-01-05                   92084                San Diego San Diego
##   vaccine_equity_metric_quartile               vem_source
## 1                              4 Healthy Places Index Score
## 2                              2 Healthy Places Index Score
## 3                              1    CDPH-Derived ZCTA Score
## 4                              4 Healthy Places Index Score
## 5                              4 Healthy Places Index Score
## 6                              2 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1               46300.3                53102                       61
## 2               22820.5                25486                       NA
## 3                 826.1                  939                       NA
## 4               32146.4                33213                       45
## 5               11136.3                12177                       NA
## 6               42677.7                47784                       12
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                           27                                0.001149
## 2                           NA                                      NA
## 3                           NA                                      NA
## 4                           30                                0.001355
## 5                           NA                                      NA
## 6                           17                                0.000251
##   percent_of_population_partially_vaccinated
## 1                                   0.000508
## 2                                         NA
## 3                                         NA
## 4                                   0.000903
## 5                                         NA
## 6                                   0.000356
##   percent_of_population_with_1_plus_dose booster_recip_count
## 1                               0.001657                  NA
## 2                                     NA                  NA
## 3                                     NA                  NA
## 4                               0.002258                  NA
## 5                                     NA                  NA
## 6                               0.000607                  NA
##                                                          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
```

There are 107 unique zip codes listed for San Diego County.

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

```
sd[which.max(sd$age12_plus_population), "zip_code_tabulation_area"]
```

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

92154

Q13. What is the overall average “Percent of Population Fully Vaccinated” value for all San Diego “County” as of “2022-02-22”?

```
# select all San Diego county entries on as_of_date == “2022-02-22”
sd.20220222 <- filter(sd, as_of_date == "2022-02-22")

# skim
skimr::skim(sd.20220222)
```

Data summary

Name	sd.20220222
Number of rows	107
Number of columns	15
Column type frequency:	
character	4
Date	1
numeric	10
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
local_health_jurisdiction	0	1	9	9	0	1	0
county	0	1	9	9	0	1	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	2022-02-22	2022-02-22	2022-02-22	1

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
zip_code_tabulation_area	0	1.00	92047.95	75.75	91901.00	92005.50	92064.00	92113.50	92173.0	▂▃▆▇▇
vaccine_equity_metric_quartile	8	0.93	2.73	1.00	1.00	2.00	3.00	4.00	4.0	▅▃▃▁▇
age12_plus_population	0	1.00	26407.70	20315.19	0.00	4305.05	26688.60	42645.80	76365.2	▇▅▃▂▂
age5_plus_population	0	1.00	28982.11	22359.43	0.00	4595.00	29040.00	46852.50	82971.0	▇▅▃▂▂
persons_fully_vaccinated	1	0.99	21890.82	17748.06	36.00	3491.25	19877.00	34445.50	77457.0	▇▅▃▂▂
persons_partially_vaccinated	1	0.99	5731.84	5551.74	18.00	982.50	4883.50	8197.00	29331.0	▇▅▂▁▁
percent_of_population_fully_vaccinated	1	0.99	0.70	0.22	0.01	0.65	0.72	0.82	1.0	▁▁▁▇▁
percent_of_population_partially_vaccinated	1	0.99	0.21	0.15	0.01	0.14	0.17	0.23	1.0	▇▃▁▁▁
percent_of_population_with_1_plus_dose	1	0.99	0.83	0.22	0.02	0.80	0.89	1.00	1.0	▁▁▁▇▁
booster_recip_count	5	0.95	8926.17	6683.90	14.00	2700.50	8947.50	13748.50	26579.0	▇▅▃▂▂

```
# find overall average
mean(sd.20220222$percent_of_population_fully_vaccinated, na.rm = TRUE) * 100
```

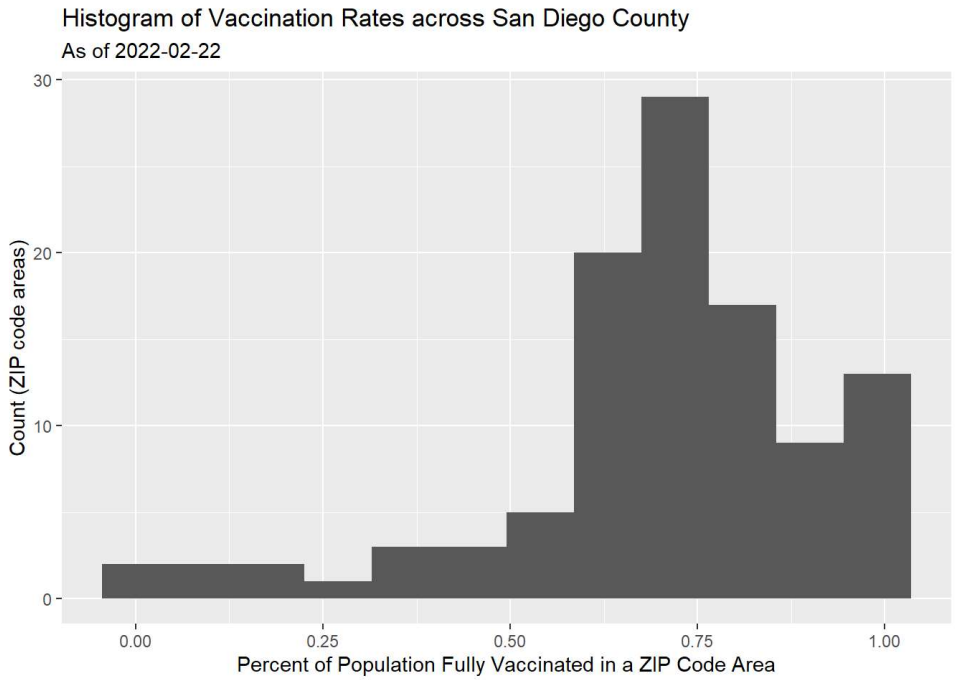
```
## [1] 70.41551
```

The overall average is 70.42%.

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 “2022-02-22”?

```
# use ggplot to make the figure
library(ggplot2)
ggplot(sd.20220222, aes(x = percent_of_population_fully_vaccinated)) +
  geom_histogram(bins = 12) +
  labs(title = "Histogram of Vaccination Rates across San Diego County", subtitle = "As of 2022-02-22", x = "Percent of Popu
lation Fully Vaccinated in a ZIP Code Area", y = "Count (ZIP code areas)")
```

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



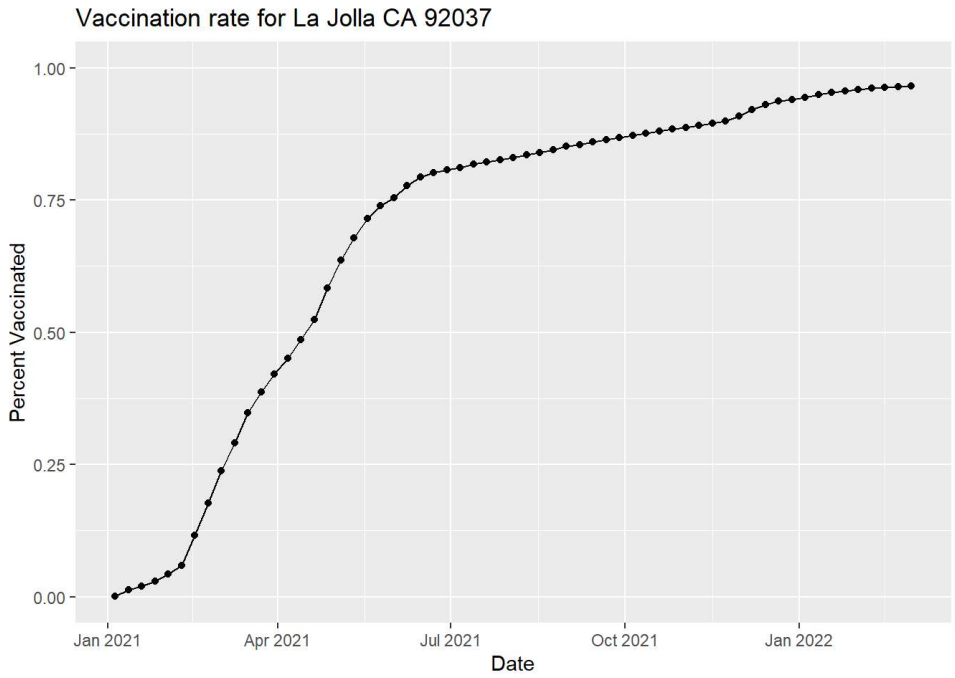
Focus on UCSD/La Jolla

```
# define selection on ucsd/La jolla area by zip code 92037 and verify population
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:

```
ggplot(ucsd) +
  aes(x = as_of_date,
      y = percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) +
  ylim(c(0,1)) +
  labs(x = "Date", y="Percent Vaccinated", title = "Vaccination rate for La Jolla CA 92037")
```



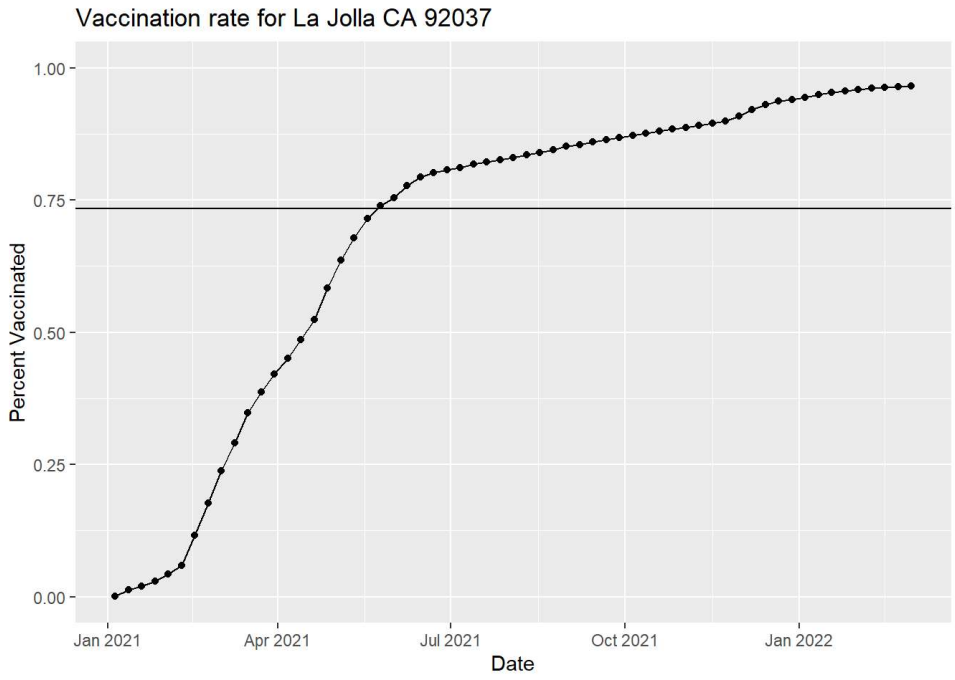
Comparing to similar sized areas

```
# Subset to all CA areas with a population as large as 92037
vax.36 <- filter(vax, age5_plus_population > 36144 &
                 as_of_date == "2022-02-22")

head(vax.36)
```

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction      county
## 1 2022-02-22                   92840                    Orange      Orange
## 2 2022-02-22                   92064                 San Diego   San Diego
## 3 2022-02-22                   92508                 Riverside   Riverside
## 4 2022-02-22                   95403                    Sonoma      Sonoma
## 5 2022-02-22                   90001               Los Angeles Los Angeles
## 6 2022-02-22                   92802                    Orange      Orange
##   vaccine_equity_metric_quartile                vem_source
## 1                              2 Healthy Places Index Score
## 2                              4 Healthy Places Index Score
## 3                              3 Healthy Places Index Score
## 4                              3 Healthy Places Index Score
## 5                              1 Healthy Places Index Score
## 6                              2 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1               47302.5                51902                    40725
## 2               42177.1                46855                    34266
## 3               32415.3                36303                    21925
## 4               38545.9                42294                    33158
## 5               47175.7                54805                    43075
## 6               35113.6                39393                    29268
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                         4324                                0.784652
## 2                         6861                                0.731320
## 3                         1714                                0.603945
## 4                         2833                                0.783988
## 5                        13917                                0.785968
## 6                         6138                                0.742975
##   percent_of_population_partially_vaccinated
## 1                                   0.083311
## 2                                   0.146430
## 3                                   0.047214
## 4                                   0.066983
## 5                                   0.253937
## 6                                   0.155814
##   percent_of_population_with_1_plus_dose booster_recip_count redacted
## 1                               0.867963               20654       No
## 2                               0.877750               15499       No
## 3                               0.651159               10753       No
## 4                               0.850971               18659       No
## 5                               1.000000               13408       No
## 6                               0.898789               12816       No
```

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 “2022-02-22”. Add this as a straight horizontal line to your plot from above with the geom_hline() function?

```
# find mean vaccination rate
ca.mean <- mean(vax.36$percent_of_population_fully_vaccinated, na.rm = TRUE)

# add this as a straight horizontal line to plot
ggplot(ucsd) +
  aes(x = as_of_date,
      y = percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) +
  ylim(c(0,1)) +
  labs(x = "Date", y="Percent Vaccinated", title = "Vaccination rate for La Jolla CA 92037") +
  geom_hline(aes(yintercept=ca.mean))
```



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 "2022-02-22"?

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

```
## [1] 0.3881090 0.6539015 0.7332750 0.8027110 1.0000000
```

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

```
## [1] 0.733385
```

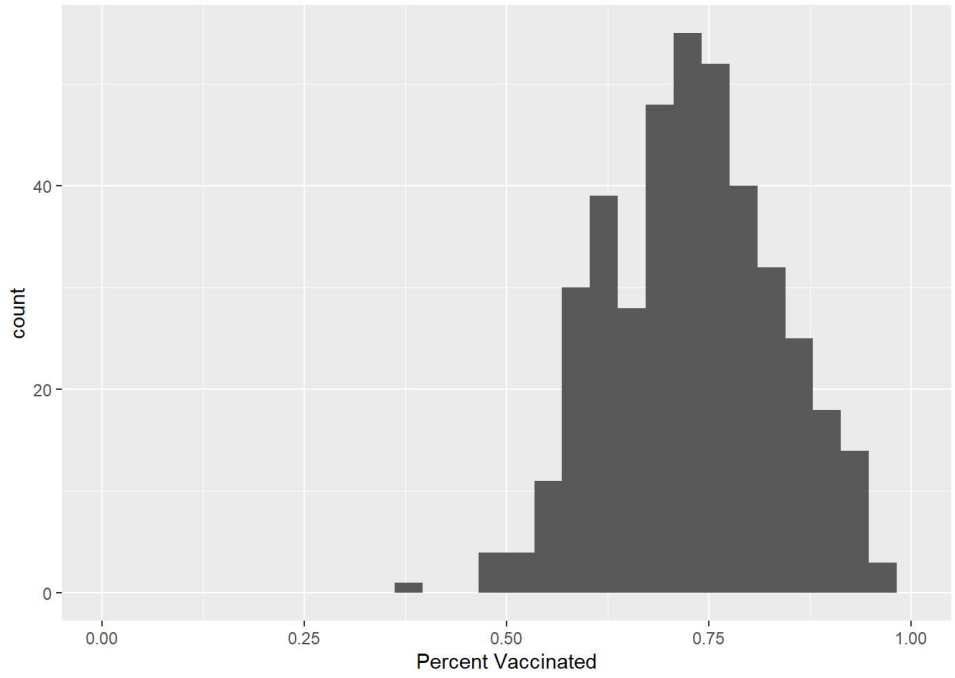
Min: 0.3881090 1st Qu.: 0.6539015 Median: 0.7332750 Mean: 0.733385 3rd Qu.: 0.8027110 Max: 1.0000000

Q18. Using ggplot generate a histogram of this data.

```
ggplot(vax.36, aes(x = percent_of_population_fully_vaccinated)) +
  geom_histogram() +
  labs(x = "Percent Vaccinated") +
  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?

```
# for 92040
vax %>% filter(as_of_date == "2022-02-22") %>%
  filter(zip_code_tabulation_area=="92040") %>%
  select(percent_of_population_fully_vaccinated)
```

```
##   percent_of_population_fully_vaccinated
## 1                               0.551304
```

```
# for 92109
vax %>% filter(as_of_date == "2022-02-22") %>%
  filter(zip_code_tabulation_area=="92109") %>%
  select(percent_of_population_fully_vaccinated)
```

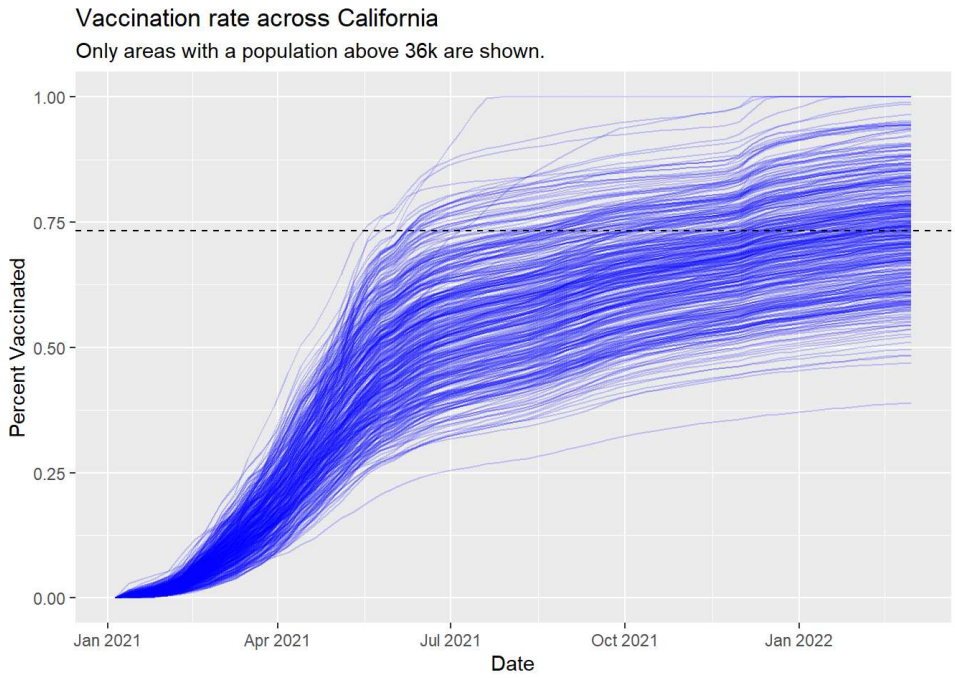
```
##   percent_of_population_fully_vaccinated
## 1                               0.723044
```

Both are below the average vlaue calculated earlier (0.55 < 0.73, 0.72 < 0.73).

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(x = 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 = ca.mean, linetype = "dashed")
```

```
## Warning: Removed 311 row(s) containing missing values (geom_path).
```



Q21. How do you feel about traveling for Spring Break and meeting for in-person class afterwards?

I'd generally feel comfortable traveling and meeting in-person in CA afterwards based on current trends in vaccination rates. Some areas might require additional caution as the (full) vaccination rate remained below 50% as of now and show no significant increase in slope.