COVID-19 Vaccination Rates

Joshua Cheung

03/02/2022

Background

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

Getting started

We move our downloaded CSV file to our project directory and then read/import into an R object called vax. We will use this data to answer the questions below.

```r
# We import the 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                    92140                 San Diego     San Diego
## 2 2021-01-05                    94133             San Francisco San Francisco
## 3 2021-01-05                    94523              Contra Costa  Contra Costa
## 4 2021-01-05                    94005                 San Mateo     San Mateo
## 5 2021-01-05                    94104             San Francisco San Francisco
## 6 2021-01-05                    94549              Contra Costa  Contra Costa
##   vaccine_equity_metric_quartile                 vem_source
## 1                             NA            No VEM Assigned
## 2                              3 Healthy Places Index Score
## 3                              4 Healthy Places Index Score
## 4                              4 Healthy Places Index Score
## 5                             NA            No VEM Assigned
## 6                              4 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1                3747.7                 3737                       NA
## 2               25070.5                25957                       NA
## 3               30457.9                32828                       NA
## 4                3996.1                 4364                       NA
## 5                 387.8                  399                       NA
## 6               25393.8                28468                       NA
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                           NA                                     NA
## 2                           NA                                     NA
## 3                           NA                                     NA
## 4                           NA                                     NA
## 5                           NA                                     NA
## 6                           NA                                     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
##                                                                 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?

Inspection of the column names shows that the column "persons_fully_vaccinated" details the total number of people fully vaccinated.

Q2. What column details the Zip code tabulation area?

Inspection of the column names shows that the column "zip_code_tabulation_area" details the zip code tabulation area.

Q3. What is the earliest date in this dataset?

```
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"
```

We see that the earliest date in this dataset is 2021-01-05.

Q4. What is the latest date in this dataset?

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

```
## [1] "2022-02-22" "2022-02-22" "2022-02-22" "2022-02-22" "2022-02-22"
## [6] "2022-02-22"
```

We see that the latest date in this dataset is 2022-02-22.

We now call the skim() function from the skimr package to get a quick overview of this dataset:

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

Table 1: Data summary

Name	vax
Number of rows	105840
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	60	0
local_health_jurisdiction	0	1	0	15	300	62	0
county	0	1	0	15	300	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	5220	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	
age12_plus_population	0	1.00	18895.04	18993.92	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	18174	0.83	12064.29	12983.91	11	1059.00	7287.50	19859.00	77213.0	
persons_partially_vaccinated	18174	0.83	820.71	1318.77	11	76.00	370.00	1066.00	31869.0	
percent_of_population_fully_vaccinated	18174	0.83	0.51	0.26	0	0.33	0.54	0.70	1.0	
percent_of_population_partially_vaccinated	18174	0.83	0.05	0.09	0	0.01	0.03	0.05	1.0	
percent_of_population_with_1_plus_dose	18174	0.83	0.54	0.27	0	0.35	0.58	0.75	1.0	
booster_recip_count	64191	0.39	3923.43	5704.10	11	169.00	1072.00	5803.00	49951.0	

Q5. How many numeric columns are in this dataset?

While the results from the skim say that there are 10 numeric columns in this dataset, we know that the column titled zip_code_tabulation_area is not technically numeric in the same way the other numeric columns are. So to answer the question there are 9 numeric column in this data set.

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] 18174
```

So there are 18174 NA values in the persons_fully_vaccinated column.

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

```
round((sum(is.na(vax$persons_fully_vaccinated))/nrow(vax))*100, digits=2)
```

```
## [1] 17.17
```

So the percentage of persons_fully_vaccinated values that are missing is 17.17%.

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

This data could be missing due to the fact that some individuals may only be partially vaccinated at the moment and have not come in to become fully vaccinated yet. Additionally, some individuals may have simply chosen to only remain partially vaccinated.

Working with dates

To start working with dates, we call the lubridate package as follows:

```
library(lubridate)
```

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

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

We can see what today's date is (at the time of writing this) as follows:

```
today()
```

```
## [1] "2022-03-07"
```

The as_of_date column of our data is currently not that usable. For example we can't easily do math with it like answering the simple question how many days have passed since data was first recorded:

```
# This will give an Error message (un-comment the following line to see this).
# today() - vax$as_of_date[1]
```

So we convert our date data into a lubridate format things like this will be much easier as well as plotting time series data later on.

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

Now we can start to do math with dates. For example: How many days have passed since the first vaccination reported in this dataset?

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

```
## Time difference of 426 days
```

Using the last and the first date value we can now determine how many days the dataset span?

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

```
## Time difference of 413 days
```

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

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

```
## Time difference of 13 days
```

So 8 days have passed since the last update of the dataset.

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

```
length(unique(as.Date(vax$as_of_date)))
```

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

SO there are 60 unique dates in the dataset.

Working with ZIP codes

We note that one of the numeric columns in the dataset (namely vax$zip_code_tabulation_area) are actually ZIP codes - a postal code used by the United States Postal Service (USPS). In R we can use the zipcodeR package to make working with these codes easier. For example, let's install in the console and then load up this package and to find the centroid of the La Jolla 92037 (i.e. UC San Diego) ZIP code area.

```
library(zipcodeR)
geocode_zip('92037')
```

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

Now we can calculate the distance between the centroids of any two ZIP codes in miles. For instance:

```
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 instance:

```
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>
```

Optional: We can use this reverse_zipcode() to pull census data later on for any or all ZIP code areas we might be interested in. For instance:

```
# We pull data for all ZIP codes in the dataset.
zipdata <- reverse_zipcode( vax$zip_code_tabulation_area )
```

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:

```
# We subset to San Diego county only areas.
sd <- vax[92109,]
```

We then use the dplyr package as follows:

```
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] 6420
```

Using dplyr is often more convenient when we are subsetting across multiple criteria - for instance, all San Diego county areas with a population of over 10,000.

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

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

```
SD <- filter(vax, county == "San Diego")
length(unique(SD$zip_code_tabulation_area))
```

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

So there are 107 distinct zip codes listed for San Diego County.

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

```
SD$zip_code_tabulation_area[which.max(SD$age12_plus_population)]
```

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

We see that 92154 is the San Diego County Zip code area that has the largest 12 + Population in this dataset.

Using dplyr we select all San Diego "county" entries on "as_of_date" "2022-02-22" and use this for the following questions.

```
sd.02 <- filter(vax, county == "San Diego" & as_of_date == "2022-02-22")
```

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

```
sd.02.fully <- sd.02$percent_of_population_fully_vaccinated
round(mean(sd.02.fully, na.rm=TRUE), digits=3)
```

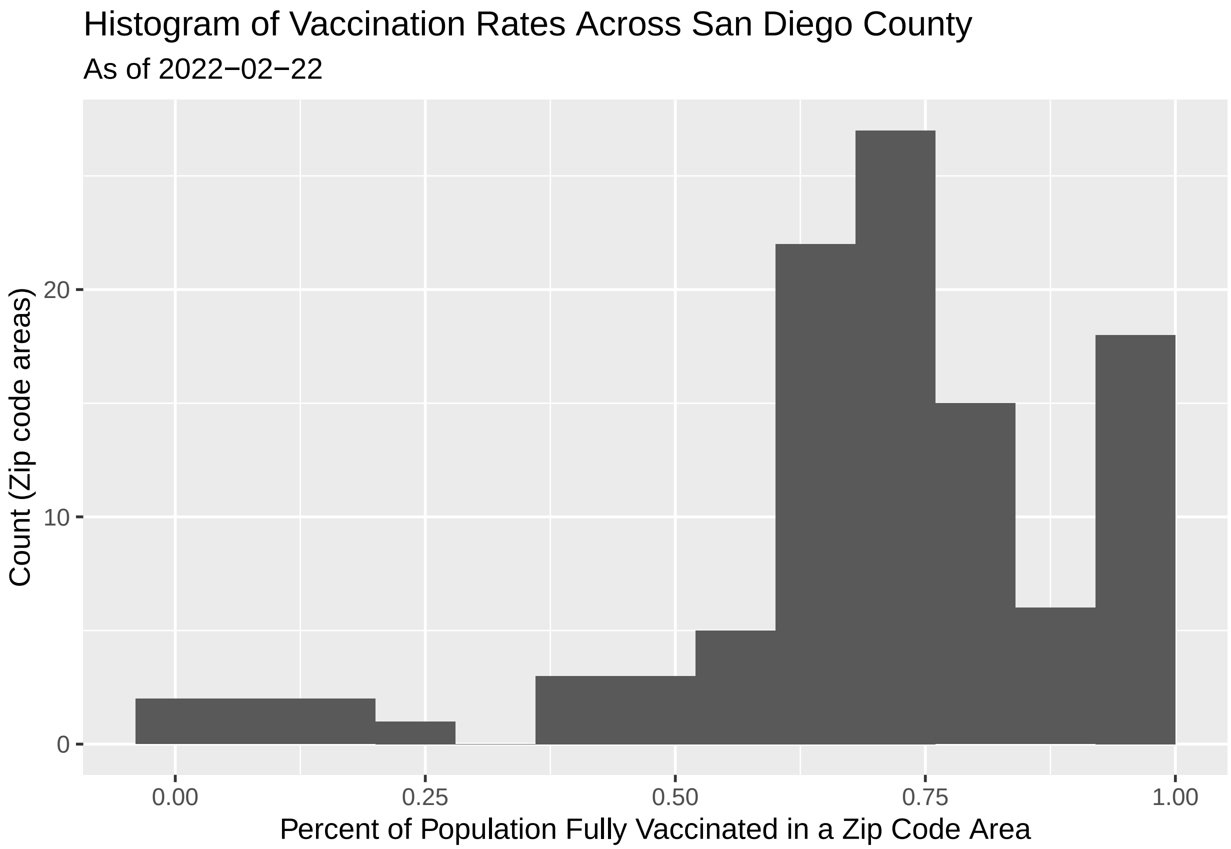
```
## [1] 0.703
```

So the overall average "Percent of Population Fully Vaccinated value for all San Diego County as of 2022-02-22 is 0.703.

> **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"?**

```
library(ggplot2)
ggplot(sd.02) +
  geom_histogram(aes(x=percent_of_population_fully_vaccinated),
                 binwidth = 0.08) +
  labs(title="Histogram of Vaccination Rates Across San Diego County",
       x="Percent of Population Fully Vaccinated in a Zip Code Area",
       y="Count (Zip code areas)",
       subtitle="As of 2022-02-22")
```

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



Focus on UCSD/La Jolla

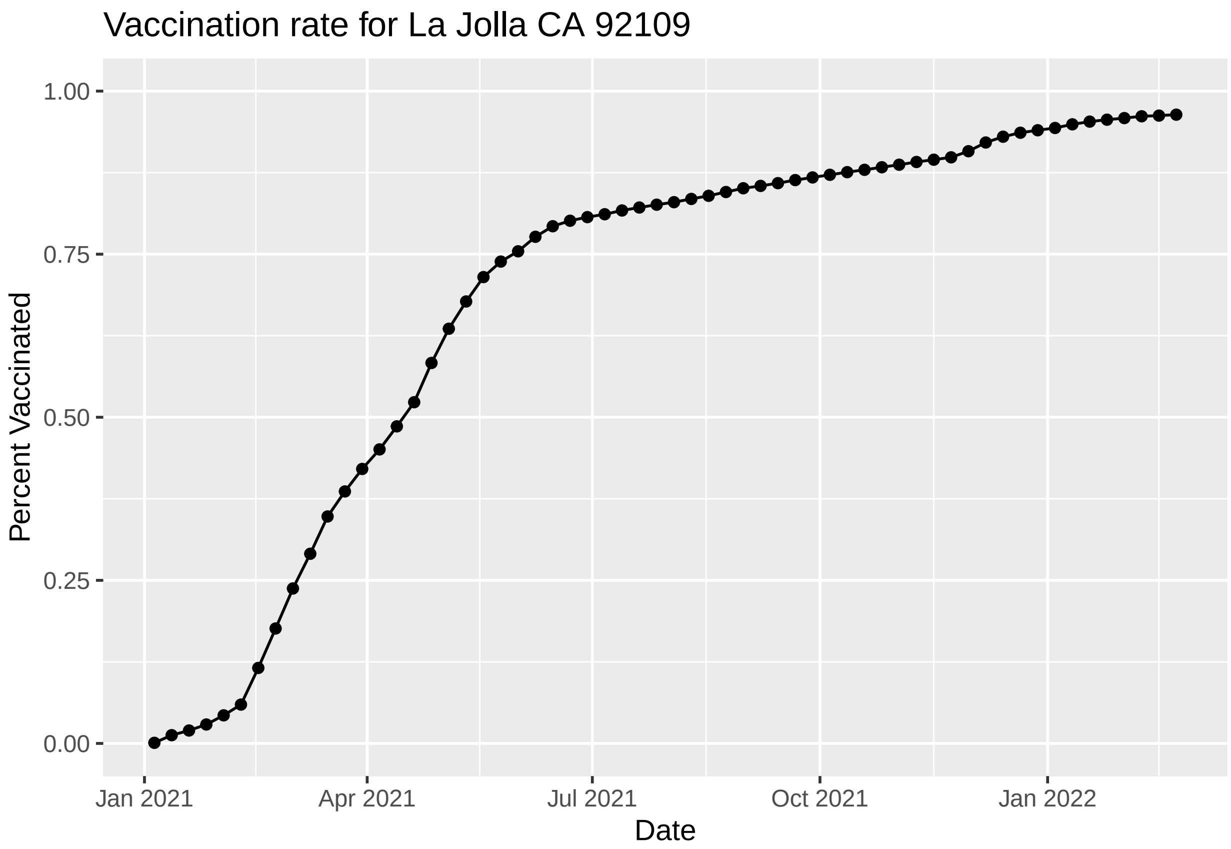
UC San Diego resides in the 92037 ZIP code area and is listed with an age 5+ population size of 36,144.

```
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(title="Vaccination rate for La Jolla CA 92109",
       x="Date", y="Percent Vaccinated")
```



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".

```
# We 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")
# We preview the first 6 rows.
head(vax.36)
```

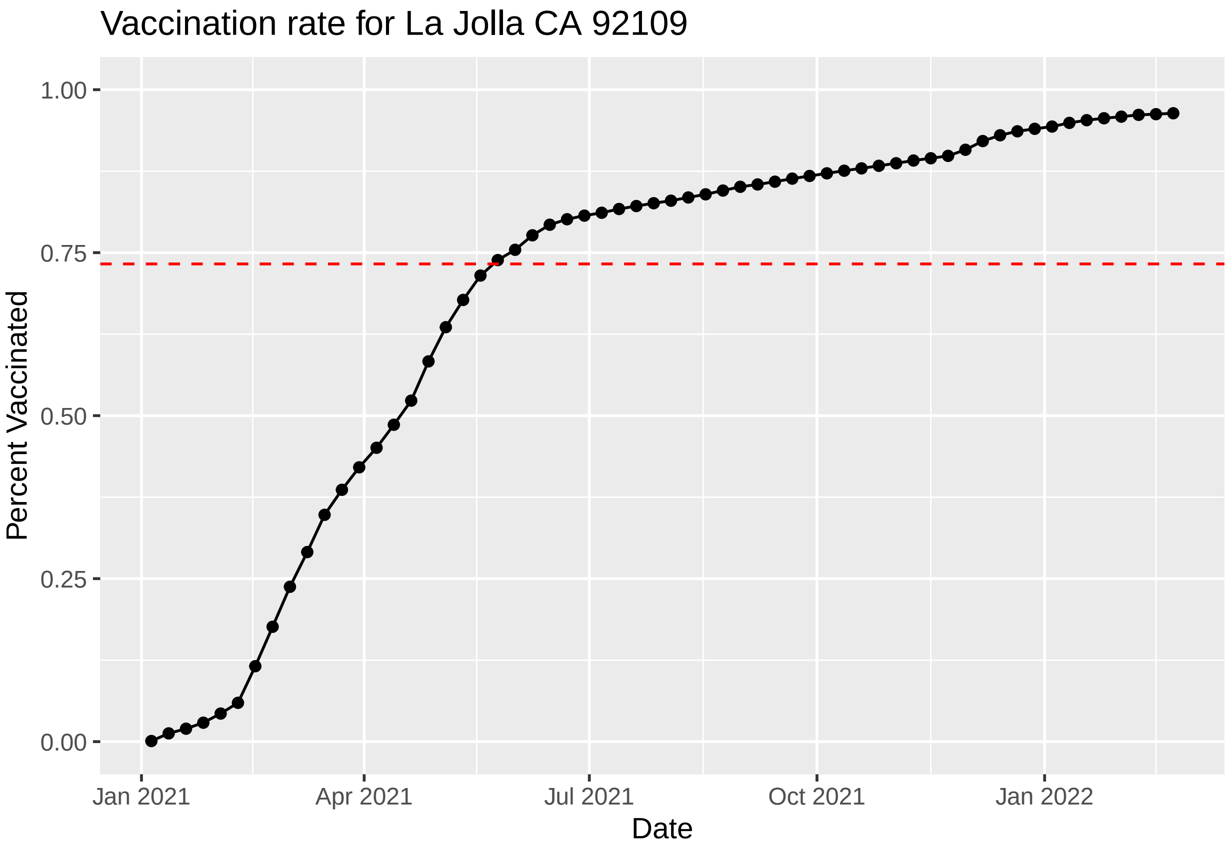
```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction       county
## 1 2022-02-22                    94582              Contra Costa Contra Costa
## 2 2022-02-22                    92592                 Riverside    Riverside
## 3 2022-02-22                    92504                 Riverside    Riverside
## 4 2022-02-22                    94546                  Alameda      Alameda
## 5 2022-02-22                    94577                  Alameda      Alameda
## 6 2022-02-22                    94565             Contra Costa Contra Costa
##   vaccine_equity_metric_quartile                 vem_source
## 1                              4 Healthy Places Index Score
## 2                              3 Healthy Places Index Score
## 3                              2 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               34809.5                40433                    42744
## 2               69581.7                79782                    44648
## 3               50996.7                56235                    32781
## 4               37839.8                41600                    37452
## 5               42041.7                45192                    39770
## 6               80663.4                90579                    74795
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                         2755                               1.000000
## 2                         5809                               0.559625
## 3                         3205                               0.582929
## 4                         3070                               0.900288
## 5                         2529                               0.880023
## 6                         5135                               0.825743
##   percent_of_population_partially_vaccinated
## 1                                   0.068137
## 2                                   0.072811
## 3                                   0.056993
## 4                                   0.073798
## 5                                   0.055961
## 6                                   0.056691
##   percent_of_population_with_1_plus_dose booster_recip_count redacted
## 1                               1.000000               27798       No
## 2                               0.632436               20599       No
## 3                               0.639922               14119       No
## 4                               0.974086               23191       No
## 5                               0.935984               24164       No
## 6                               0.882434               36596       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?**

```
# The mean is as follows:
mean <- mean(vax.36$percent_of_population_fully_vaccinated)
mean
```

```
## [1] 0.732736
```

```
ggplot(ucsd) +
  aes(x=as_of_date,
      y=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") +
  geom_hline(yintercept=mean, linetype="dashed", color = "red")
```



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"?

```
# We already computed the mean earlier.
# The five number summary can be found as follows:
fivenum(vax.36$percent_of_population_fully_vaccinated)
```

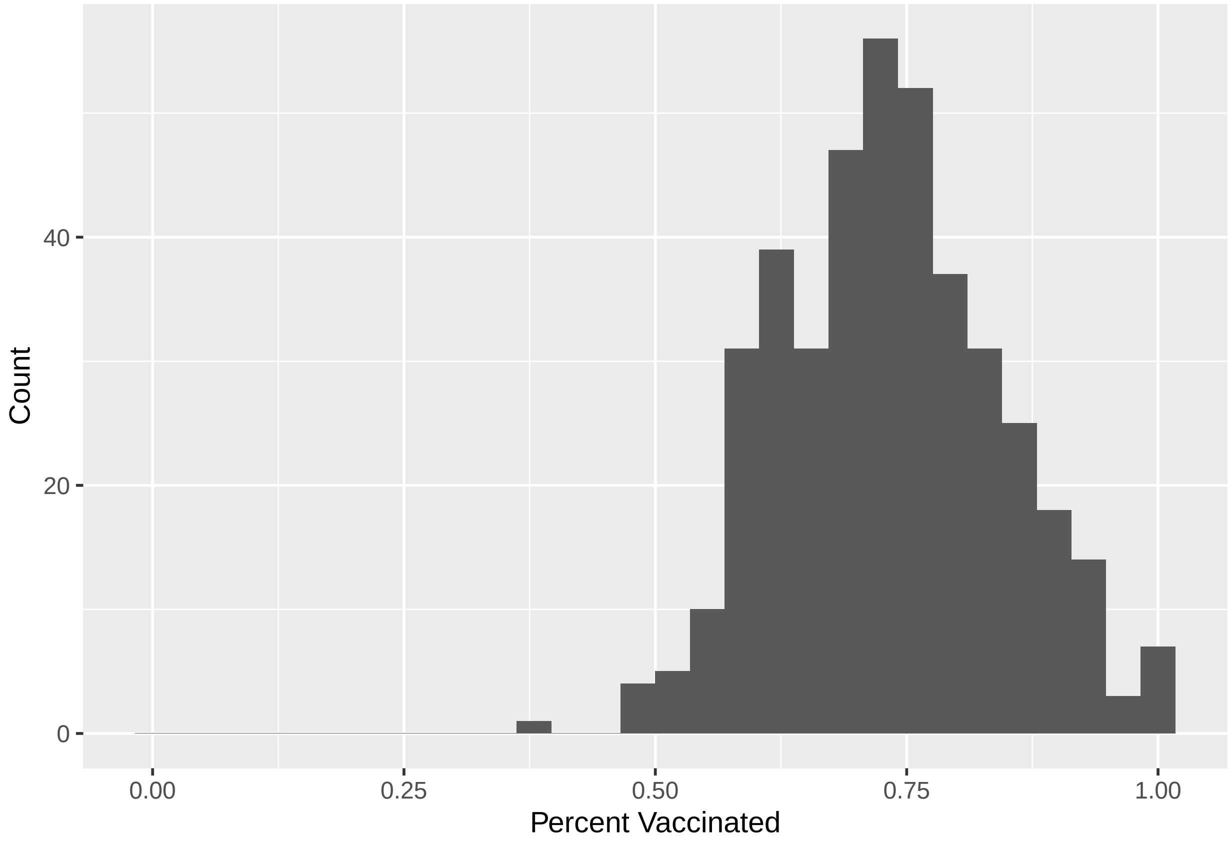
```
## [1] 0.3878320 0.6533895 0.7326670 0.8024260 1.0000000
```

Thus for the 6 number summary, the minimum is 0.3878320, the 1st quartile is 0.6533895, the median is 0.7326670, the mean is 0.732736, the 3rd quartile is 0.8024260, and the maximum is 1.0000000.

Q18. Using ggplot generate a histogram of this data.

```
ggplot(vax.36) +
  geom_histogram(aes(x=percent_of_population_fully_vaccinated)) +
  labs(x="Percent Vaccinated",
       y="Count") +
  expand_limits(x = 0, y = 0)
```

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



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

```
##   percent_of_population_fully_vaccinated
## 1                                0.55093
```

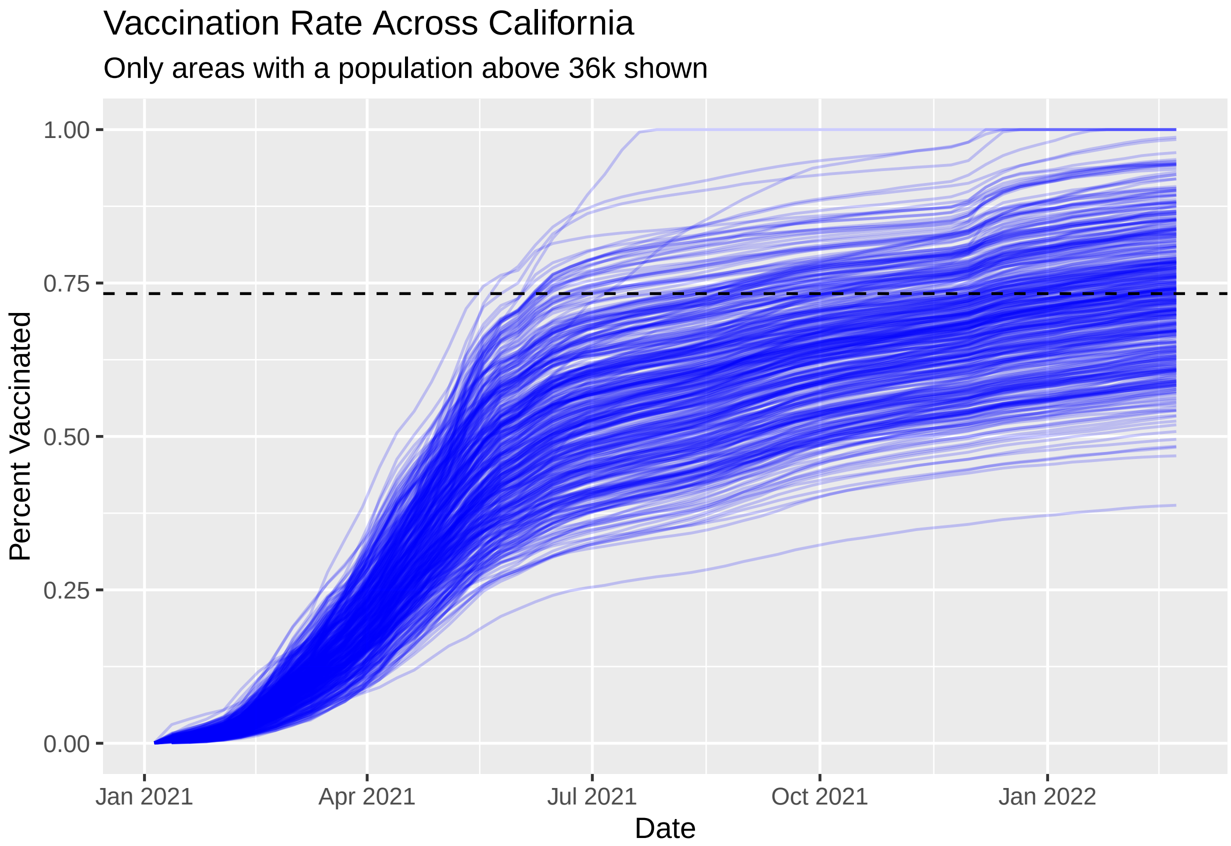
The 92109 and 92040 ZIP code areas are below the average value we calculated.

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

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


ggplot(vax.36.all) +
  aes(x=as_of_date,
      y=percent_of_population_fully_vaccinated,
      group=zip_code_tabulation_area) +
  geom_line(alpha=0.2, color="blue") +
  ylim(0,1.00) +
  labs(x="Date", y="Percent Vaccinated",
       title="Vaccination Rate Across California",
       subtitle="Only areas with a population above 36k shown") +
  geom_hline(yintercept = mean, linetype="dashed")
```

```
## Warning: Removed 309 row(s) containing missing values (geom_path).
```



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

While the fact that the percent of the population vaccinated against COVID-19 is generally trending upwards is certainly encouraging, the data we analyzed is only for California. While travelling within California may be less of a risk, it is possible that travelling outside of California still poses some unknown risk as we do now know the vaccination rates in other states. Additionally, even if one remains within California over the break, others who have traveled to states with low vaccination rates could pose a health risk to other if they were to attend in-person classes. Thus travelling over Spring Break and meeting for in-person classes afterwards could still potentially pose some risks.