BIMM 143 Class 17

Anika Bhattacharjya (A15459876)

11/23/2021

Getting Started

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

Q1. What column details the total number of people fully vaccinated?

The 9th column that says "persons_fully_vaccinated."

Q2. What column details the Zip code tabulation area?

The 2nd column labeled "zip_code_tabulation_data."

```
library(lubridate)
```

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

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

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

Q3. What is the earliest date in this dataset?

```
vax$as_of_date[1]
```

```
## [1] "2021-01-05"
```

Q4. What is the latest date in this dataset?

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

```
## [1] "2021-11-16"
```

```
library(devtools)
```

```
## Loading required package: usethis
```

```
library(skimr)
```

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

Table 1: Data summary

Name	vax
Number of rows	81144

Table 1: Data summary

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_1_plus_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?

9

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)?

```
library(plyr)
```

```
signif(sum( is.na(vax$persons_fully_vaccinated) )/81144*100,2)
```

```
## [1] 10
```

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

It may be missing because of human record keeping error or no one reporting vaccinations.

Working with dates

```
library(lubridate)
```

```
today()
```

```
## [1] "2021-11-27"
```

Look at the as_of_date column

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

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

```
## Time difference of 326 days
```

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

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

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

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

```
## Time difference of 11 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
```

Working with Zip Codes

"zipcodeR" wouldn't work for me so Professor Grant said to skip this section.

Focus on San Diego Area

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

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

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

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

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

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

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

```
## [1] 23
```

```
sd$zip_code_tabulation_area[23]
```

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

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

```
sd.vax <- filter(vax, county == "San Diego" &
              as_of_date == "2021-11-09")
```

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

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

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"?

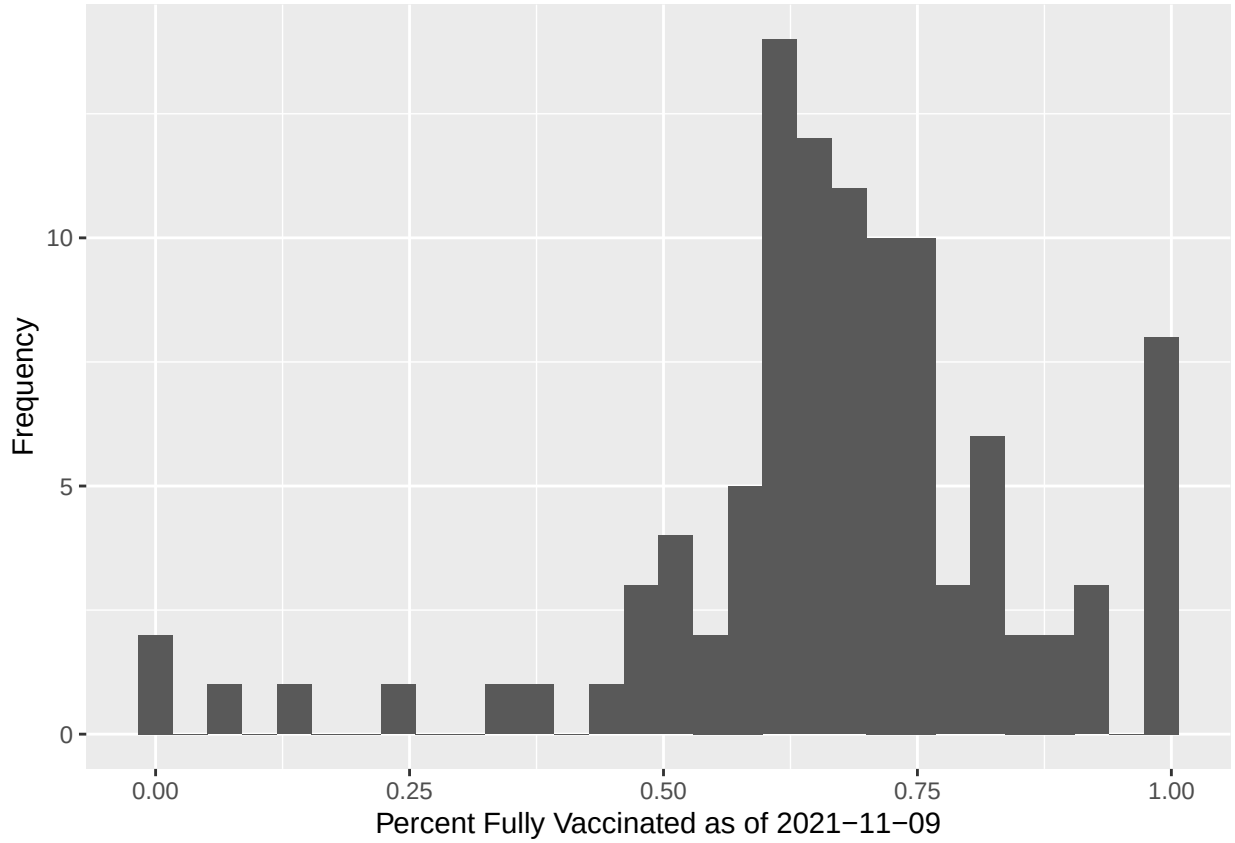
```
library(ggplot2)
```

```
?ggplot
```

```
ggplot(sd.vax) + geom_histogram(aes(x=percent_of_population_fully_vaccinated)) + labs(x = "Percent Fully
```

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

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



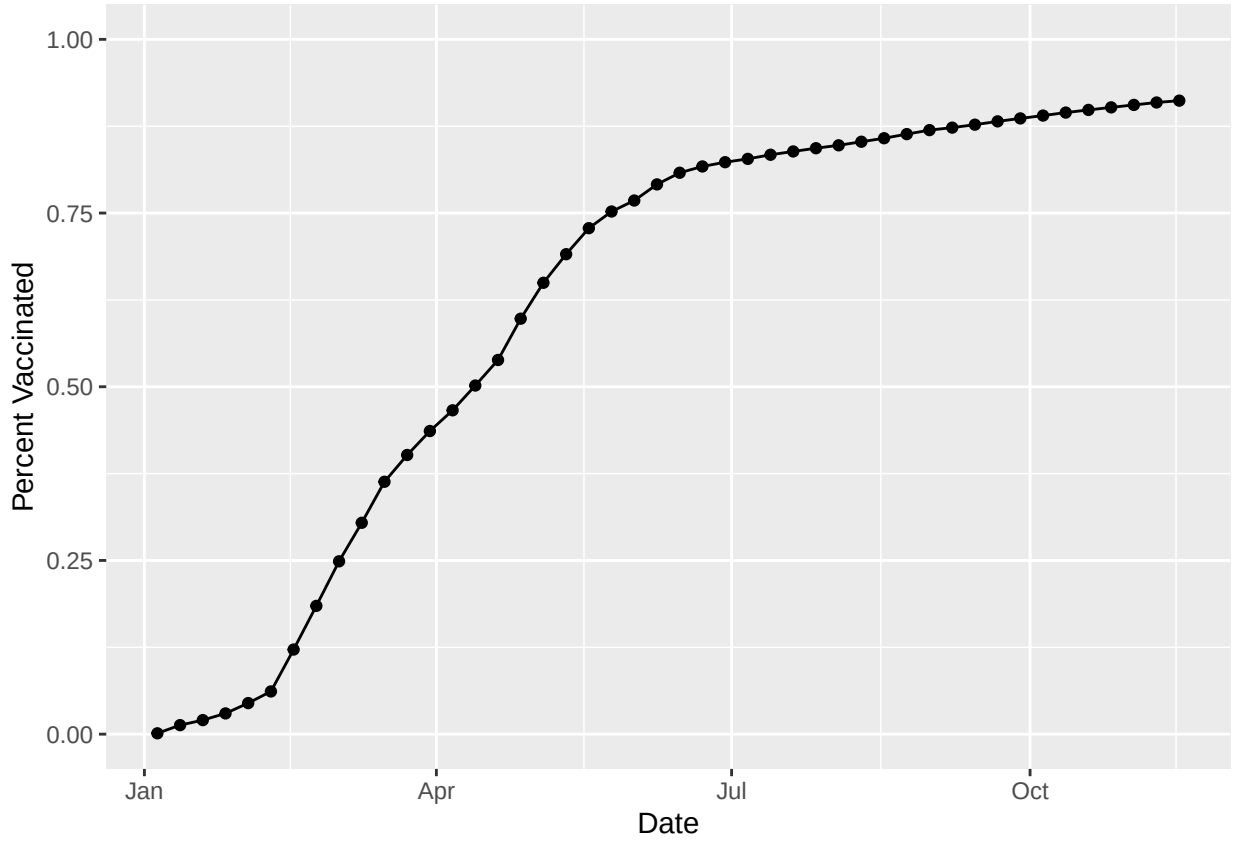
Focus on UCSD/La Jolla

```
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

```
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")
```

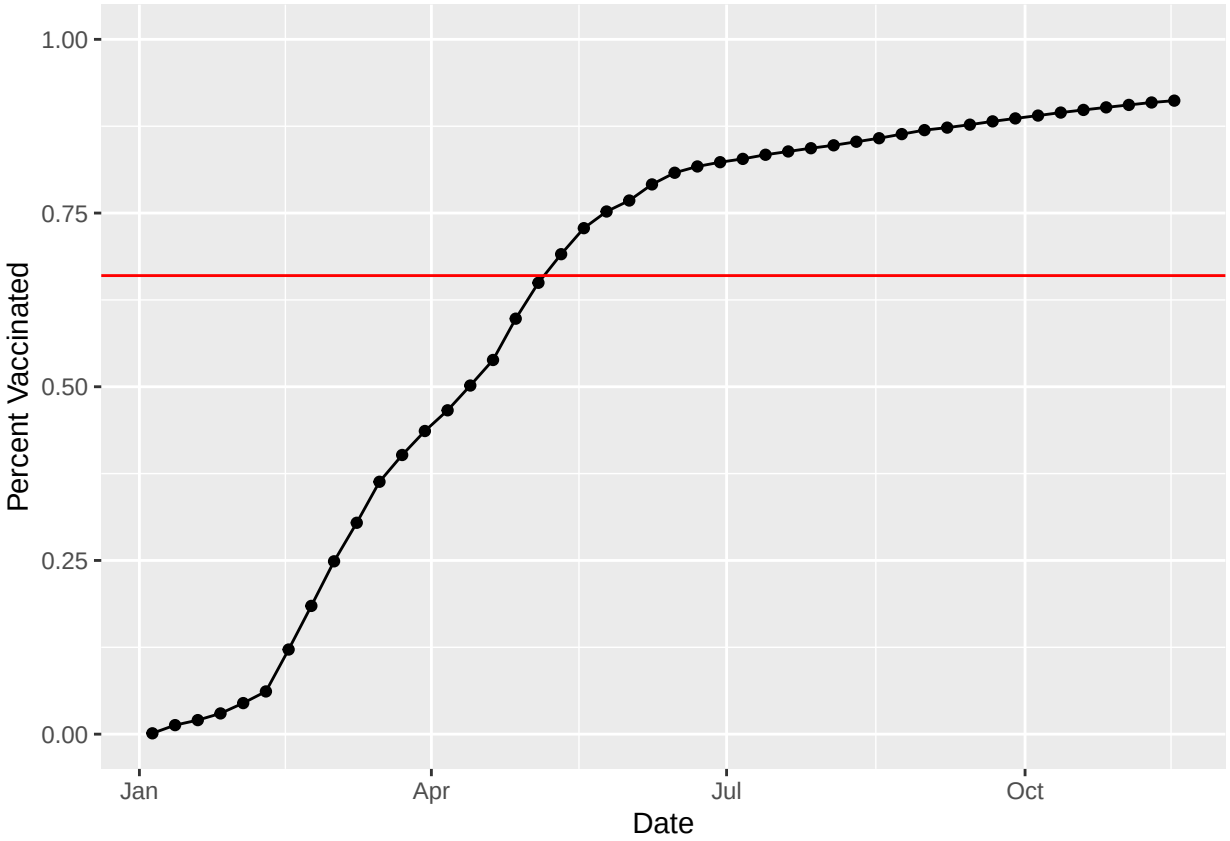


Comparing 92037 to other similar sized areas

```
# Subset to all CA areas with a population as large as 92037
vax.36 <- filter(vax, age5_plus_population > 36144 &
                 as_of_date == "2021-11-16")
```

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?

```
ggplot(ucsd) +
  aes(as_of_date,
      percent_of_population_fully_vaccinated) +
  geom_point() +
  geom_line(group=1) + geom_hline(yintercept = 0.66, col = "red") +
  ylim(c(0,1)) +
  labs(x = "Date", y="Percent Vaccinated")
```



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"?

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

```
##        0%       25%       50%       75%      100%
## 0.3518830 0.5890990 0.6648930 0.7286045 1.0000000
```

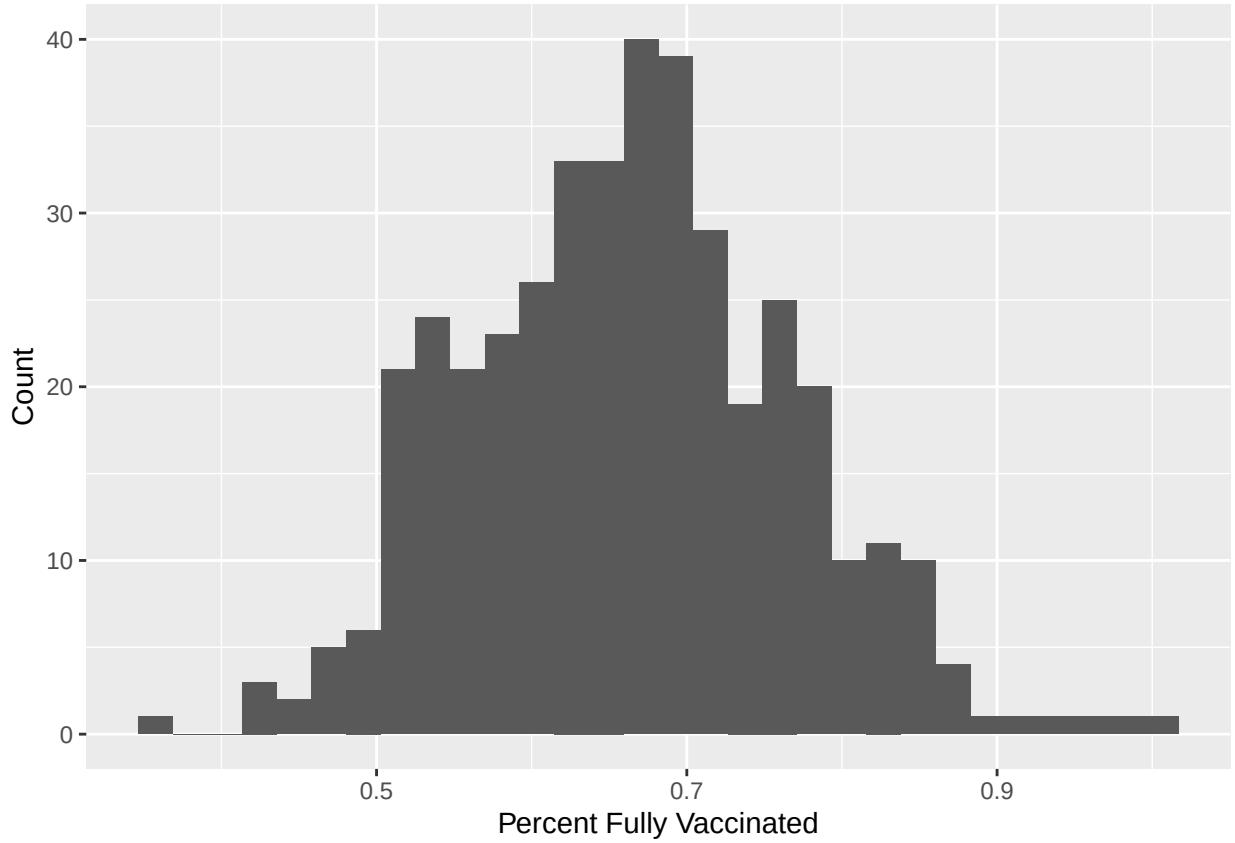
```
mean(vax.36$percent_of_population_fully_vaccinated)
```

```
## [1] 0.6629812
```

Q18. Using ggplot generate a histogram of this data

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

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

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

```
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
## 1                               0.687763
```

92040 is below average (0.6629812) and 92109 is above.

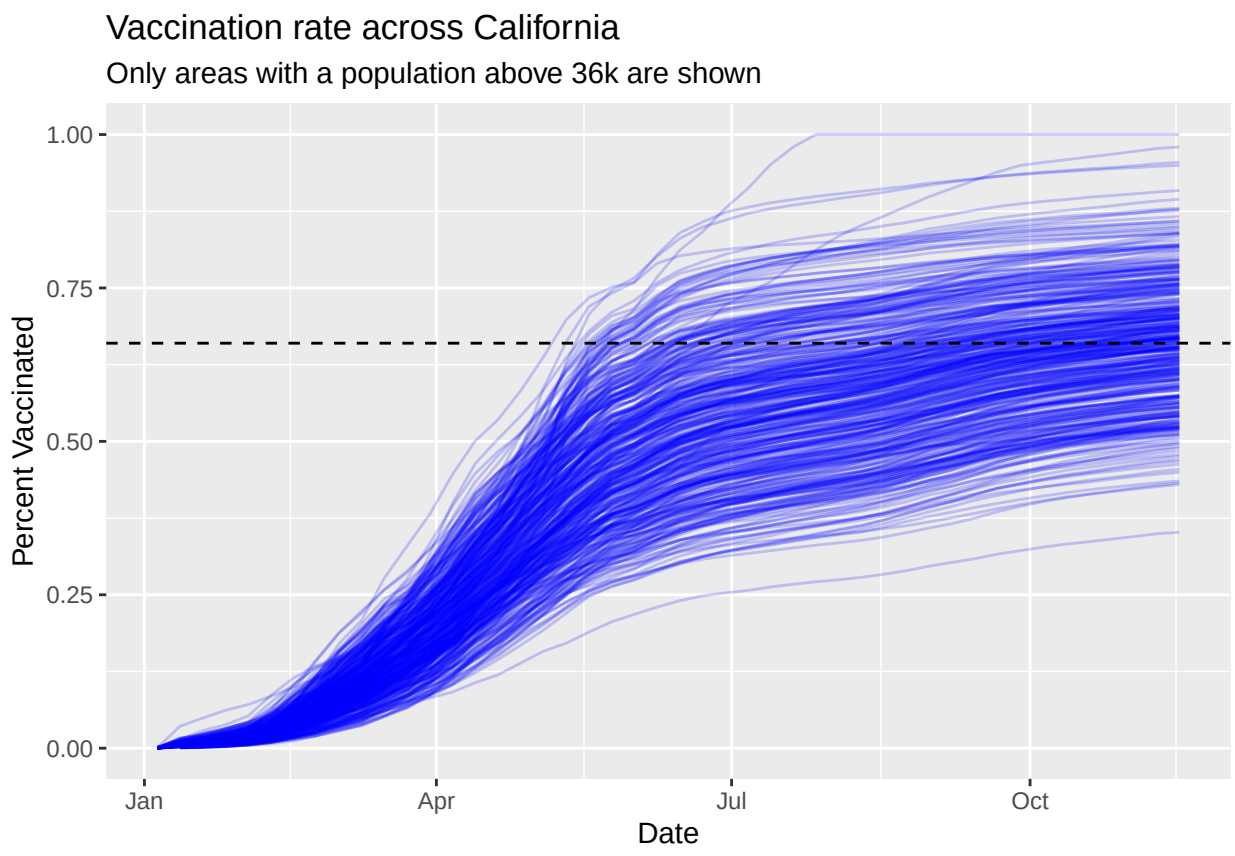
Q20. Finally make a time course plot of vaccination progress for all areas in the full dataset with a age5_plus_population > 36144.

```
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 = 0.66, linetype= "dashed")
```

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



Q21. How do you feel about traveling for Thanksgiving and meeting for in-person class next Week?

I think it would be better to do it virtually since people won't have time to get properly tested by Tuesday if they come back on Sunday.