Class17 Mini Project

Stefanie Hodapp (PID: A53300084)

12/3/2021

1. Background

```
# 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                    92091                 San Diego    San Diego
## 2 2021-01-05                    92116                 San Diego    San Diego
## 3 2021-01-05                    95360                Stanislaus   Stanislaus
## 4 2021-01-05                    94564              Contra Costa Contra Costa
## 5 2021-01-05                    95501                  Humboldt     Humboldt
## 6 2021-01-05                    95492                    Sonoma       Sonoma
##   vaccine_equity_metric_quartile                 vem_source
## 1                              4    CDPH-Derived ZCTA Score
## 2                              3 Healthy Places Index Score
## 3                              1 Healthy Places Index Score
## 4                              4 Healthy Places Index Score
## 5                              2 Healthy Places Index Score
## 6                              4 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1                1238.3                 1303                       NA
## 2               30255.7                31673                       45
## 3               10478.5                12301                       NA
## 4               17033.0                18381                       NA
## 5               20566.6                22061                       NA
## 6               25076.9                28024                       NA
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                           NA                                     NA
## 2                          898                               0.001421
## 3                           NA                                     NA
## 4                           NA                                     NA
## 5                           NA                                     NA
## 6                           NA                                     NA
##   percent_of_population_partially_vaccinated
## 1                                         NA
## 2                                   0.028352
## 3                                         NA
## 4                                         NA
## 5                                         NA
## 6                                         NA
##   percent_of_population_with_1_plus_dose
## 1                                     NA
## 2                               0.029773
## 3                                     NA
## 4                                     NA
## 5                                     NA
## 6                                     NA
##                                                                 redacted
## 1 Information redacted in accordance with CA state privacy requirements
## 2                                                                     No
## 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
```

2. Getting Started

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

```
which(colnames(vax)=="persons_fully_vaccinated") # Column 9
```

```
## [1] 9
```

Q2. What column details the Zip code tabulation area?

```
which(colnames(vax)=="zip_code_tabulation_area" ) # Column 2
```

```
## [1] 2
```

Q3. What is the earliest date in this dataset?

```
min(vax$as_of_date) # 2021-01-05
```

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

Q4. What is the latest date in this dataset?

```
max(vax$as_of_date) # 2021-11-30
```

```
## [1] "2021-11-30"
```

Q5. How many numeric columns are in this dataset?

```
# call the skim() function from the skimr package to get a quick overview of this dataset
skimr::skim(vax)
```

Table 1: Data summary

Name	vax
Number of rows	84672
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	48	0
local_health_jurisdiction	0	1	0	15	240	62	0
county	0	1	0	15	240	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	4176	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	8472	0.90	9709.47	11714.06	11	526.00	4309.50	16316.00	71552.0	
persons_partially_vaccinated	8472	0.90	1891.41	2100.88	11	197.00	1268.50	2874.00	20158.0	
percent_of_population_fully_vaccinated	8472	0.90	0.43	0.27	0	0.21	0.45	0.63	1.0	
percent_of_population_partially_vaccinated	8472	0.90	0.10	0.10	0	0.06	0.07	0.11	1.0	
percent_of_population_with_1_plus_dose	8472	0.90	0.51	0.26	0	0.31	0.54	0.71	1.0	

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)) # 8472
```

```
## [1] 8472
```

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

```
signif((sum(is.na(vax$persons_fully_vaccinated)) / nrow(vax))*100, 2) # 10%
```

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

2.1 Working with Dates

```
library(lubridate)
```

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

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

```
today()
```

```
## [1] "2021-12-04"
```

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

# Compute the number of days that have passed since the first vaccination reported in this dataset
today() - vax$as_of_date[1]
```

```
## Time difference of 333 days
```

```
# Determine how many days the dataset span
vax$as_of_date[nrow(vax)] - vax$as_of_date[1]
```

```
## Time difference of 329 days
```

> Q9. How many days have passed since the last update of the dataset? Time difference of 332 days

> Q10. How many unique dates are in the dataset (i.e. how many different dates are detailed)? Time difference of 329 days

3. Working with ZIP codes

```
# library(zipcodeR)
```

```
# find the centroid of the La Jolla 92037 (i.e. UC San Diego) ZIP code area
#geocode_zip('92037')

# Calculate the distance between the centroids of any two ZIP codes in miles, e.g.
# zip_distance('92037','92109')

# pull census data about ZIP code areas
# reverse_zipcode(c('92037', "92109") )
```

3.1 Focus on the San Diego area

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

```
## [1] 5136
```

```r
# subset all San Diego county areas with a population of over 10,000
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
sd.10 <- filter(vax, county == "San Diego" &
                age5_plus_population > 10000)
```

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

```r
length(unique(sd$zip_code_tabulation_area)) # 107 unique zip codes
```

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

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

```r
sd$zip_code_tabulation_area[which.max(sd$age12_plus_population)] # 92154
```

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

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

```r
# Filter data by as_of_date == 2021-11-09
sd_nov_9 <- filter(sd, as_of_date == "2021-11-09")
head(sd_nov_9)
```

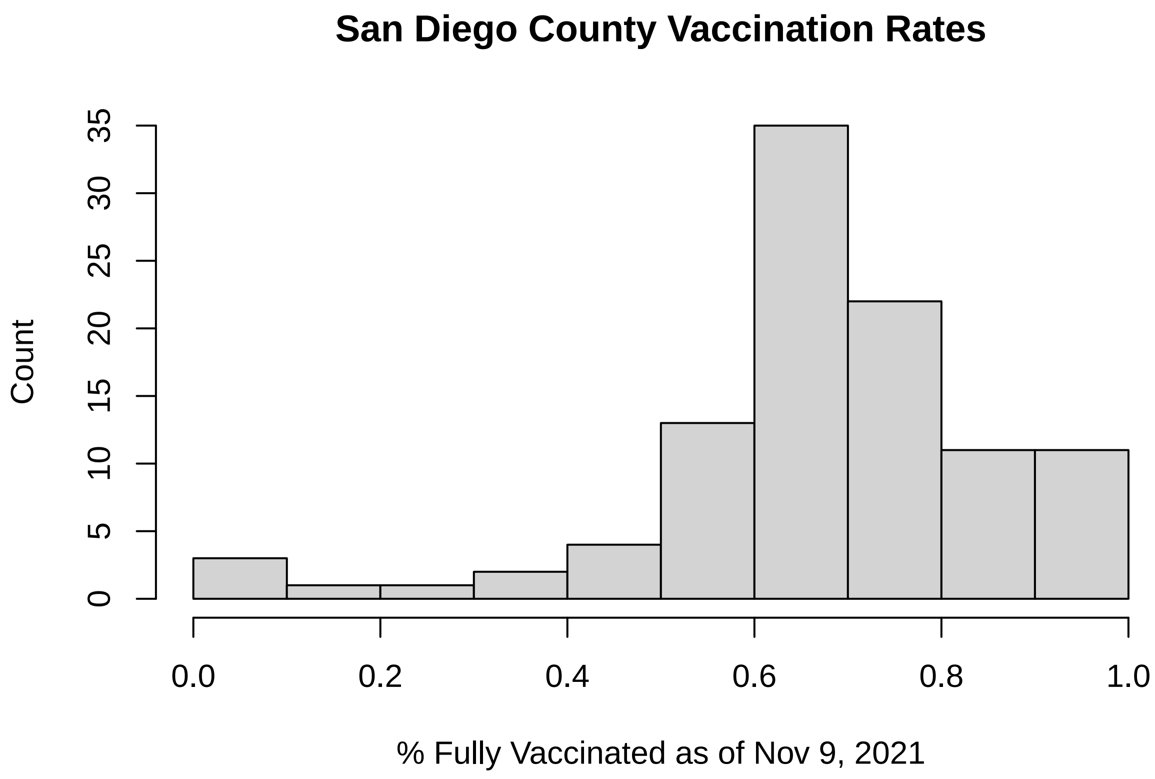
```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 1 2021-11-09                    91978                 San Diego San Diego
## 2 2021-11-09                    92069                 San Diego San Diego
## 3 2021-11-09                    91942                 San Diego San Diego
## 4 2021-11-09                    91917                 San Diego San Diego
## 5 2021-11-09                    92126                 San Diego San Diego
## 6 2021-11-09                    92154                 San Diego San Diego
##   vaccine_equity_metric_quartile                 vem_source
## 1                              2 Healthy Places Index Score
## 2                              2 Healthy Places Index Score
## 3                              3 Healthy Places Index Score
## 4                              1    CDPH-Derived ZCTA Score
## 5                              4 Healthy Places Index Score
## 6                              2 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1                8644.9                 9663                     6028
## 2               41447.3                46850                    29661
## 3               34685.9                37483                    25410
## 4                 826.1                  939                      877
## 5               71820.2                77775                    54229
## 6               76365.2                82971                    69195
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                          761                               0.623823
## 2                         3691                               0.633106
## 3                         3234                               0.677907
## 4                          162                               0.933972
## 5                         5943                               0.697255
## 6                        11026                               0.833966
##   percent_of_population_partially_vaccinated
## 1                                   0.078754
## 2                                   0.078783
## 3                                   0.086279
## 4                                   0.172524
## 5                                   0.076413
## 6                                   0.132890
##   percent_of_population_with_1_plus_dose redacted
## 1                               0.702577       No
## 2                               0.711889       No
## 3                               0.764186       No
## 4                               1.000000       No
## 5                               0.773668       No
## 6                               0.966856       No
```

```
# Calculate the average of percent of population fully vaccinated for all of san diego county
mean(sd_nov_9$percent_of_population_fully_vaccinated, na.rm=TRUE)*100 # 67.40809%
```

```
## [1] 67.40809
```

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"?

```
# histogram using base R
hist(sd_nov_9$percent_of_population_fully_vaccinated, main ="San Diego County Vaccination Rates", xlab =
```



```
# histogram using ggplot
# ggplot(data=sd_nov_9, aes(x=percent_of_population_fully_vaccinated)) + geom_histogram() + labs(title=
```

3.1.1 Focus on UCSD/La Jolla

```
# Filter data for UC San Diego (UCSD resides in the 92037 ZIP code area and is listed with an age 5+ po
ucsd <- filter(sd, zip_code_tabulation_area=="92037")
ucsd[1,]$age5_plus_population
```

```
## [1] 36144
```

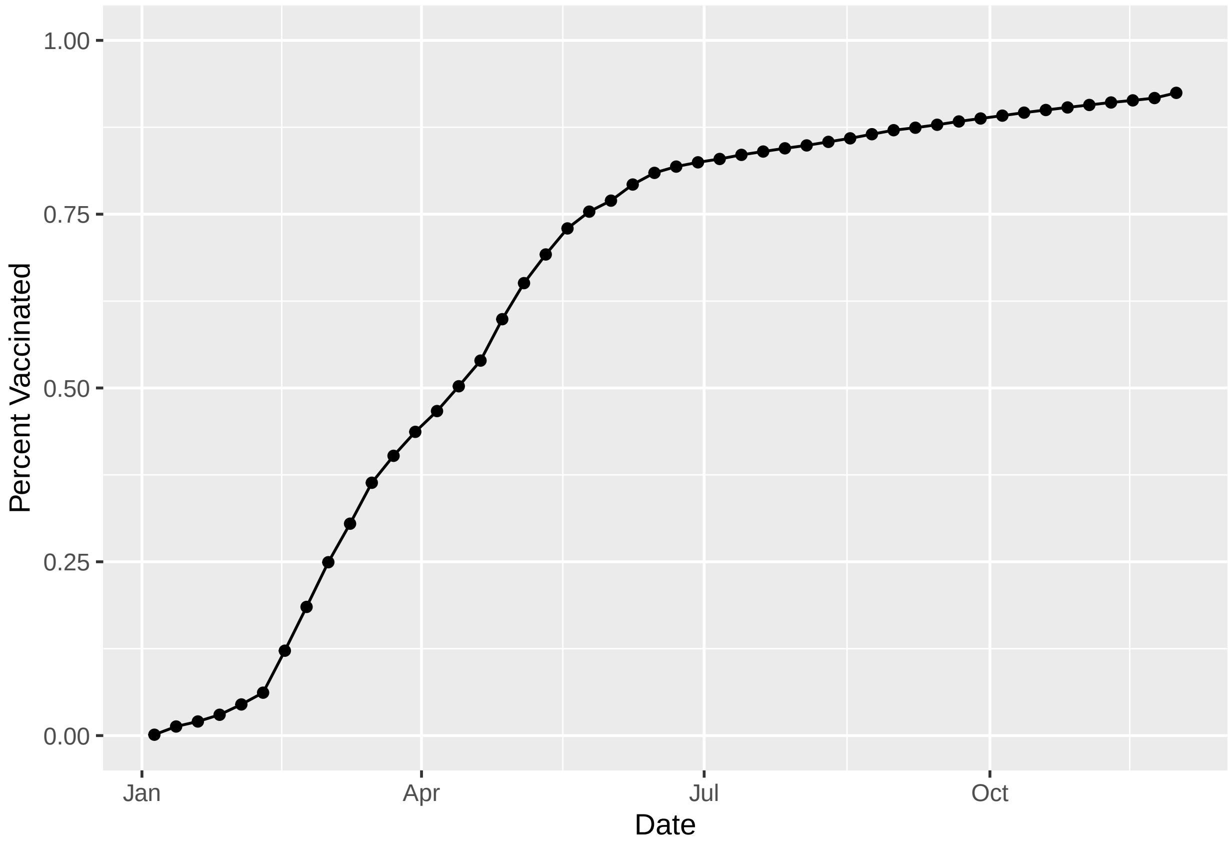
```
head(ucsd)
```

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction    county
## 1 2021-01-05                    92037                 San Diego San Diego
## 2 2021-01-12                    92037                 San Diego San Diego
## 3 2021-01-19                    92037                 San Diego San Diego
## 4 2021-01-26                    92037                 San Diego San Diego
## 5 2021-02-02                    92037                 San Diego San Diego
## 6 2021-02-09                    92037                 San Diego San Diego
##   vaccine_equity_metric_quartile                 vem_source
## 1                              4 Healthy Places Index Score
## 2                              4 Healthy Places Index Score
## 3                              4 Healthy Places Index Score
## 4                              4 Healthy Places Index Score
## 5                              4 Healthy Places Index Score
## 6                              4 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1               33675.6                36144                       46
## 2               33675.6                36144                      473
## 3               33675.6                36144                      734
## 4               33675.6                36144                     1083
## 5               33675.6                36144                     1620
## 6               33675.6                36144                     2232
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                         1270                               0.001273
## 2                         1572                               0.013087
## 3                         3518                               0.020308
## 4                         6220                               0.029963
## 5                         8416                               0.044821
## 6                         9663                               0.061753
##   percent_of_population_partially_vaccinated
## 1                                   0.035137
## 2                                   0.043493
## 3                                   0.097333
## 4                                   0.172089
## 5                                   0.232846
## 6                                   0.267347
##   percent_of_population_with_1_plus_dose redacted
## 1                               0.036410       No
## 2                               0.056580       No
## 3                               0.117641       No
## 4                               0.202052       No
## 5                               0.277667       No
## 6                               0.329100       No
```

Q15. Using ggplot make a graph of the vaccination rate time course for the 92037 ZIP code area:

```r
library(ggplot2)

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")
```



3.1.2 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 == "2021-11-16")

head(vax.36)
```

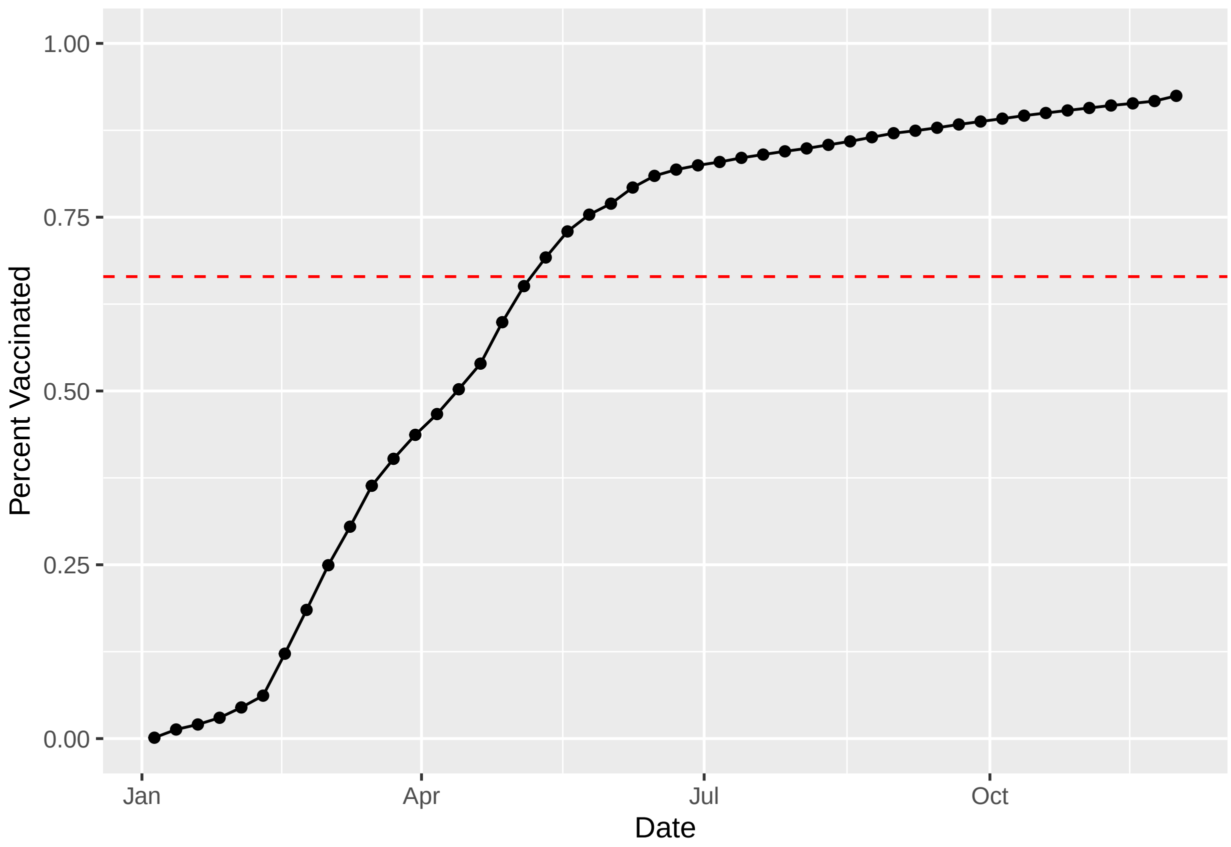
```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction         county
## 1 2021-11-16                    92345            San Bernardino San Bernardino
## 2 2021-11-16                    92553                 Riverside      Riverside
## 3 2021-11-16                    92058                 San Diego      San Diego
## 4 2021-11-16                    91786            San Bernardino San Bernardino
## 5 2021-11-16                    92507                 Riverside      Riverside
## 6 2021-11-16                    93021                   Ventura        Ventura
##   vaccine_equity_metric_quartile                 vem_source
## 1                              1 Healthy Places Index Score
## 2                              1 Healthy Places Index Score
## 3                              1 Healthy Places Index Score
## 4                              2 Healthy Places Index Score
## 5                              1 Healthy Places Index Score
## 6                              4 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1               66047.5               75539                    35432
## 2               61770.8               70472                    37411
## 3               34956.0               39695                    14023
## 4               45602.3               50410                    30834
## 5               51432.5               55253                    31939
## 6               32753.7               36197                    24918
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                         4389                               0.469056
## 2                         4846                               0.530863
## 3                         2589                               0.353269
## 4                         3132                               0.611664
## 5                         3427                               0.578050
## 6                         2012                               0.688400
##   percent_of_population_partially_vaccinated
## 1                                   0.058102
## 2                                   0.068765
## 3                                   0.065222
## 4                                   0.062131
## 5                                   0.062024
## 6                                   0.055585
##   percent_of_population_with_1_plus_dose redacted
## 1                               0.527158       No
## 2                               0.599628       No
## 3                               0.418491       No
## 4                               0.673795       No
## 5                               0.640074       No
## 6                               0.743985       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 "2021-11-16". Add this as a straight horizontal line to your plot from above with the geom_hline() function?

```
# Calculate mean "Percent of Population Fully Vaccinated" for ZIP code areas with a population as large
mean_pop <- mean(vax.36$percent_of_population_fully_vaccinated)
mean_pop
```

```
## [1] 0.6645132
```

```
# Add mean_pop as a straight horizontal line to your plot from above with the geom_hline() function
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") +
  geom_hline(yintercept=mean_pop, linetype = 'dashed', col = 'red')
```



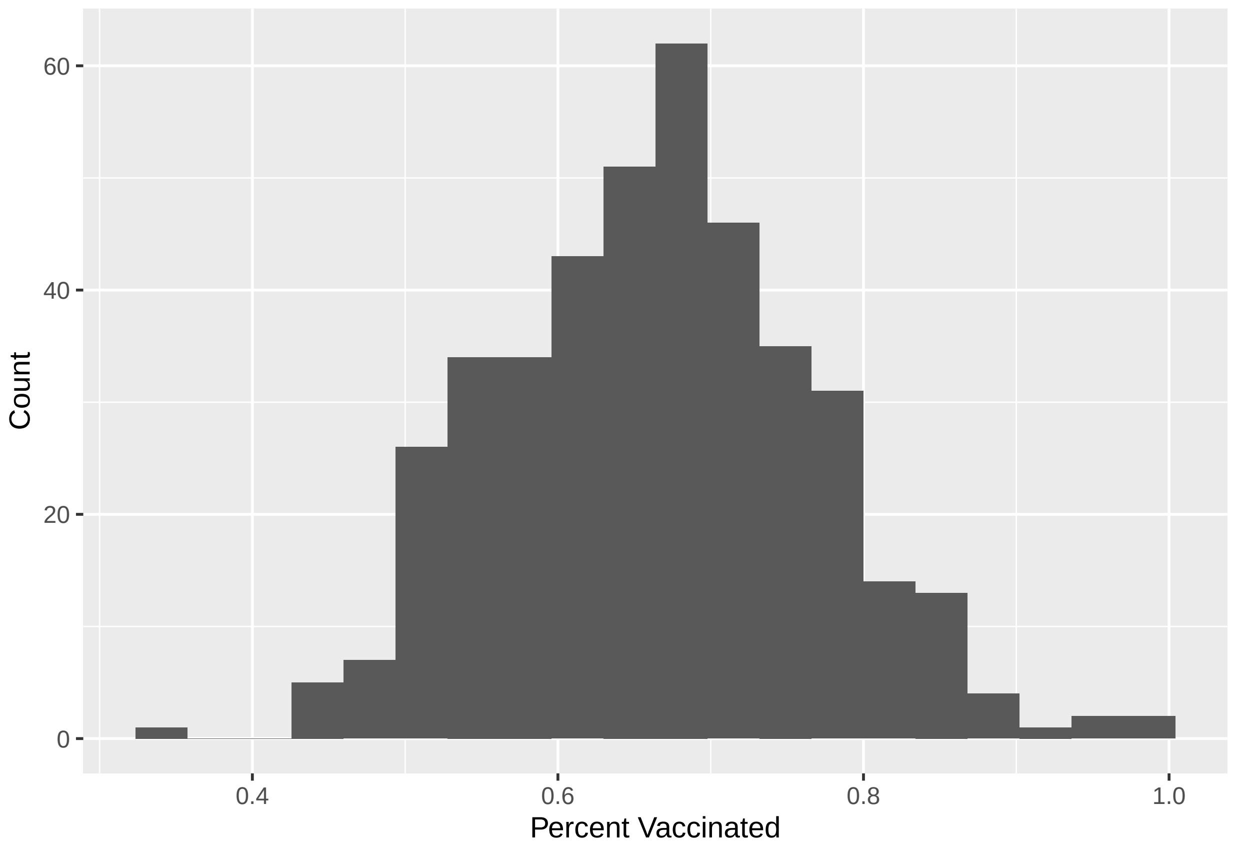
Q17. What is the 6 number summary (Min, 1st Qu., Median, Mean, 3rd Qu., and Max) of the "Percent of Population Fully Vaccinated" values for ZIP code areas with a population as large as 92037 (La Jolla) as_of_date "2021-11-16"?

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.3533  0.5910  0.6669  0.6645  0.7311  1.0000
```

Q18. Using ggplot generate a histogram of this data.

```
ggplot(vax.36) +
  aes(x=percent_of_population_fully_vaccinated) +
  geom_histogram(bins=20) +
  labs(x= "Percent Vaccinated", y="Count")
```



Q19. Is the 92109 and 92040 ZIP code areas above or below the average value you calculated for all these above? The 92040 zip code is below the average value ($0.52142 < 0.6645132$) while the 92109 zip code is above the average value ($0.68912 > 0.6645132$).

```
# 92040
vax %>% filter(as_of_date == "2021-11-16") %>%
  filter(zip_code_tabulation_area=="92040") %>%
  select(percent_of_population_fully_vaccinated)   # 0.52142
```

```
##   percent_of_population_fully_vaccinated
## 1                                0.52142
```

```
# 92109
vax %>% filter(as_of_date == "2021-11-16") %>%
  filter(zip_code_tabulation_area=="92109") %>%
  select(percent_of_population_fully_vaccinated)   # 0.68912
```

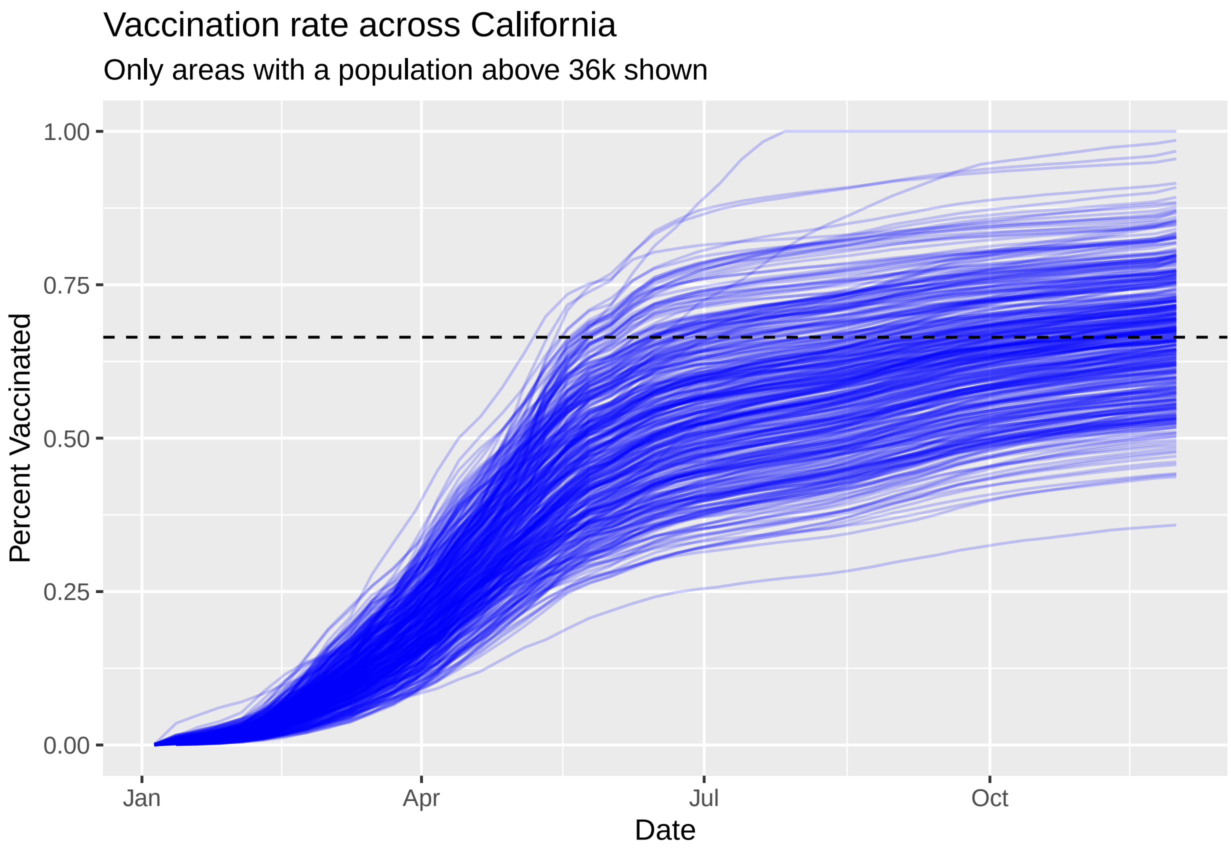
```
##   percent_of_population_fully_vaccinated
## 1                                0.68912
```

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(c(0,1)) +
  labs(x="Date", y="Percent Vaccinated",
       title="Vaccination rate across California",
       subtitle="Only areas with a population above 36k shown") +
  geom_hline(yintercept = mean_pop, linetype="dashed")
```

```
## Warning: Removed 177 row(s) containing missing values (geom_path).
```



Q21. How do you feel about traveling for Thanksgiving and meeting for in-person class next Week? N/A, but I am glad we had our last class in-person! Thank you for an excellent quarter!