

# Class17

Leyna Nguyen (PID A15422197)

11/23/2021

## 1. Getting Started

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

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction   county
## 1 2021-01-05                92804                Orange    Orange
## 2 2021-01-05                92626                Orange    Orange
## 3 2021-01-05                92250            Imperial    Imperial
## 4 2021-01-05                92637                Orange    Orange
## 5 2021-01-05                92155            San Diego    San Diego
## 6 2021-01-05                92259            Imperial    Imperial
##   vaccine_equity_metric_quartile          vem_source
## 1                             2 Healthy Places Index Score
## 2                             3 Healthy Places Index Score
## 3                             1 Healthy Places Index Score
## 4                             3 Healthy Places Index Score
## 5                             NA          No VEM Assigned
## 6                             1    CDPH-Derived ZCTA Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1                76455.9                84200                19
## 2                44238.8                47883                NA
## 3                 7098.5                 8026                NA
## 4                16027.4                16053                NA
## 5                 456.0                 456                NA
## 6                 119.0                 121                NA
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                        1282                        0.000226
## 2                         NA                        NA
## 3                         NA                        NA
## 4                         NA                        NA
## 5                         NA                        NA
## 6                         NA                        NA
##   percent_of_population_partially_vaccinated
## 1                        0.015226
## 2                         NA
## 3                         NA
## 4                         NA
## 5                         NA
```

```
## 6 NA
## percent_of_population_with_1_plus_dose
## 1 0.015452
## 2 NA
## 3 NA
## 4 NA
## 5 NA
## 6 NA
## redacted
## 1 No
## 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
```

```
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] "2021-11-16" "2021-11-16" "2021-11-16" "2021-11-16" "2021-11-16"
## [6] "2021-11-16"
```

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

The column called “persons\_fully\_vaccinated” details the total number of people fully vaccinated.

Q2. What column details the Zip code tabulation area?

The column called “zip\_code\_tabulation\_area” details the zip code tabulation area.

Q3. What is the earliest date in this dataset?

The earliest date is 2021-01-05.

Q4. What is the latest date in this dataset?

The latest date is 2021-11-16.

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

Table 1: Data summary

Name	vax

Number of rows	81144
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	46	0
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: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
zip_code_tabulation_area	0	1.00	93665.111817.39	90001	92257.7593658.5095380.5097635.0					
vaccine_equity_metric_quartile	0	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	
age12_plus_population	0	1.00	18895.0418993.94	0	1346.95	13685.1031756.1288556.7				
age5_plus_population	0	1.00	20875.2421106.05	0	1460.50	15364.0034877.00101902.0				
persons_fully_vaccinated	8256	0.90	9456.49	11498.25	11	506.00	4105.00	15859.0071078.0		
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_8256plus_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?

```
ncol(vax)
```

```
## [1] 14
```

There are 9 numeric columns.

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
```

There are 8256 NA values in the persons\_fully\_vaccinated column.

Q7. What percent of persons\_fully\_vaccinated values are missing (to 2 significant figures)?

```
sum(vax$persons_fully_vaccinated)
```

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

We will use the **lubridate** package to make life a lot easier when dealing with dates and times:

```
library(lubridate)
```

```
##
```

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

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

```
##
```

```
##      date, intersect, setdiff, union
```

```
today()
```

```
## [1] "2021-11-23"
```

Now I can do useful math with dates more easily:

```
# Specify that we are using the Year-month-day format
```

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

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

```
## Time difference of 322 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 7 days
```

7 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(vax$as_of_date))
```

```
## [1] 46
```

```
46*7
```

```
## [1] 322
```

There are 46 unique dates in the dataset.

```
library(zipcodeR)
```

```
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
```

```
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>
```

## Pull data for all ZIP codes in the dataset

```
#zipdata <- reverse_zipcode( vax$zip_code_tabulation_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: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
```

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?

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

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction   county
## 23 2021-01-05           92154                San Diego San Diego
##   vaccine_equity_metric_quartile                vem_source
## 23                2 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 23           76365.2           82971                32
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 23                1336                0.000386
##   percent_of_population_partially_vaccinated
## 23                0.016102
##   percent_of_population_with_1_plus_dose redacted
## 23                0.016488                No
```

The zip code area 92154 has the largest 12+ population in this dataset.

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

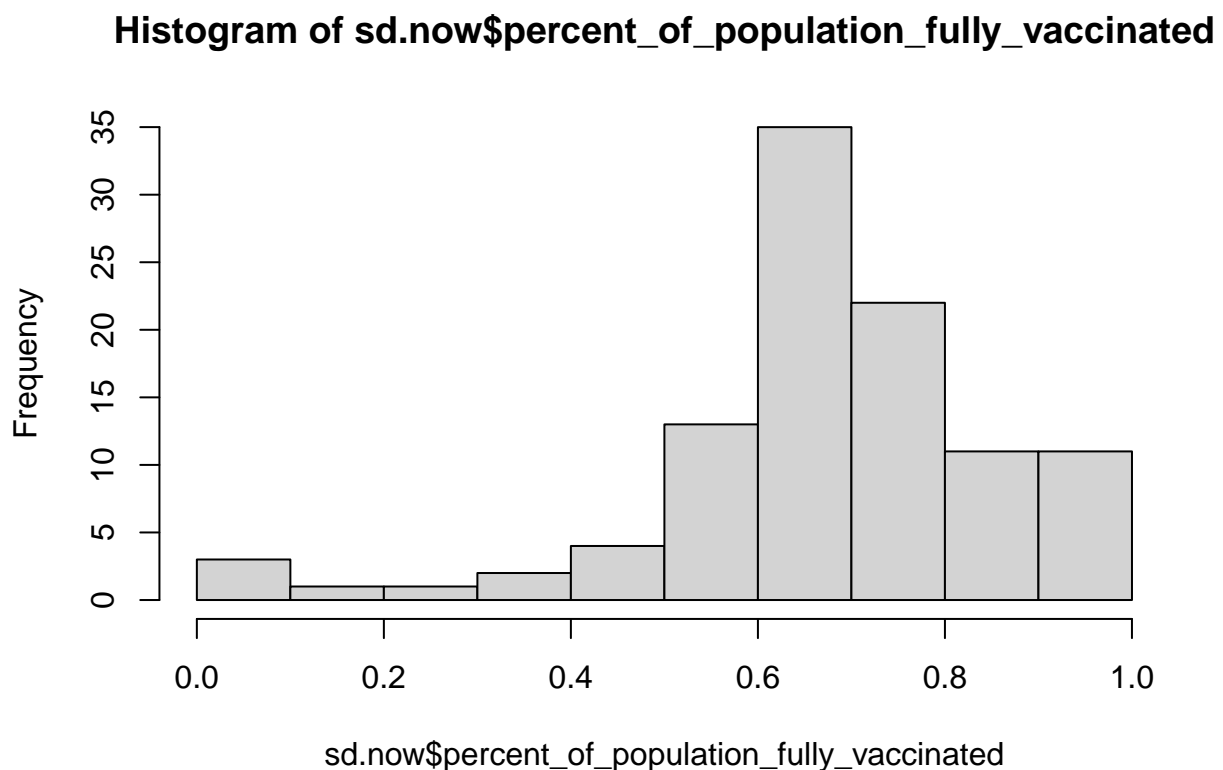
```
sd.now <- filter(sd, as_of_date=="2021-11-09")
mean(sd.now$percent_of_population_fully_vaccinated, na.rm=TRUE)
```

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

The overall average is 67.3%.

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”?

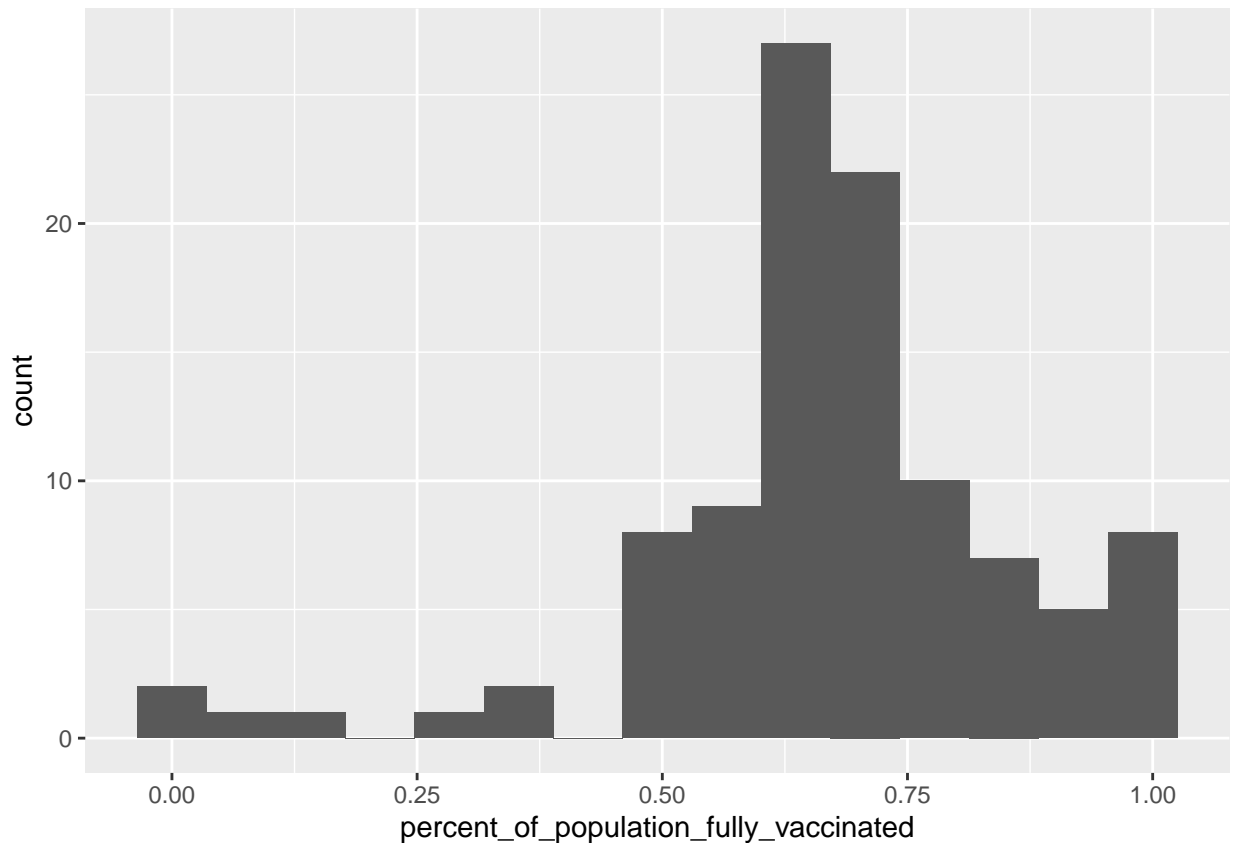
```
hist(sd.now$percent_of_population_fully_vaccinated)
```



```
library(ggplot2)

ggplot(sd.now) +
  aes(percent_of_population_fully_vaccinated) +
  geom_histogram(bins=15)
```

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



```
ucsd <- filter(sd, zip_code_tabulation_area=="92037")
ucsd[1,]$age5_plus_population
```

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

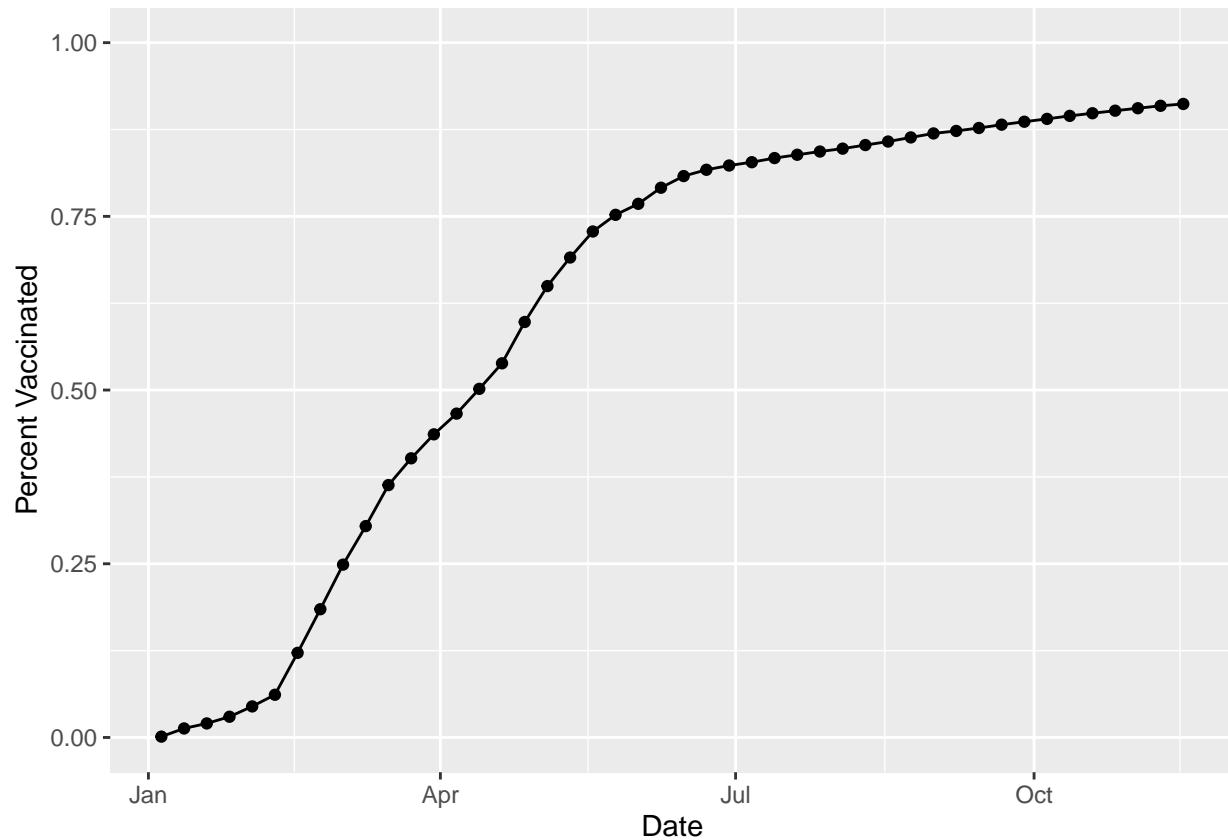
Q15. Using ggplot make a graph of the vaccination rate time course for the 92037 ZIP code area:

```
filter(sd.now, zip_code_tabulation_area=="92037")
```

```
##   as_of_date zip_code_tabulation_area local_health_jurisdiction   county
## 1 2021-11-09                92037                San Diego San Diego
##   vaccine_equity_metric_quartile                vem_source
## 1                        4 Healthy Places Index Score
##   age12_plus_population age5_plus_population persons_fully_vaccinated
## 1             33675.6             36144             32859
##   persons_partially_vaccinated percent_of_population_fully_vaccinated
## 1                6354                                0.909114
##   percent_of_population_partially_vaccinated
## 1                                0.175797
##   percent_of_population_with_1_plus_dose redacted
## 1                                1           No
```



```
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")
```



```
# Subset to all CA areas with a population as large as 92037
ucsd[1,]$age5_plus_population
```

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

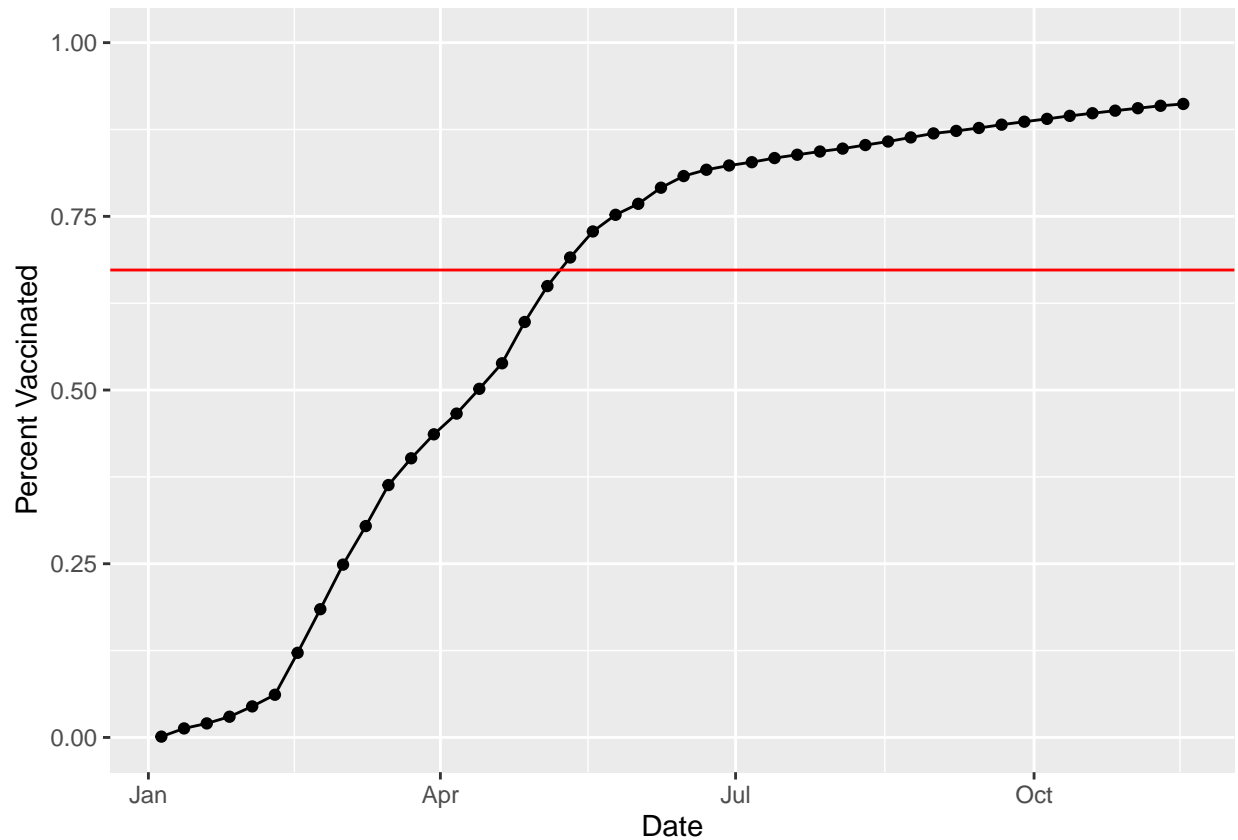
```
vax.36 <- filter(vax, age5_plus_population > 36144 &
  as_of_date == "2021-11-16")
```

```
#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 “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=0.67276, color="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”?

```
summary(vax.36)
```

```
##   as_of_date      zip_code_tabulation_area local_health_jurisdiction
## Min.   :2021-11-16 Min.   :90001                      Length:411
## 1st Qu.:2021-11-16 1st Qu.:91762                      Class :character
## Median :2021-11-16 Median :92646                      Mode  :character
## Mean   :2021-11-16 Mean   :92862
## 3rd Qu.:2021-11-16 3rd Qu.:94517
## Max.   :2021-11-16 Max.   :96003
##   county      vaccine_equity_metric_quartile vem_source
## Length:411    Min.   :1.000                      Length:411
## Class :character 1st Qu.:1.000                      Class :character
## Mode  :character Median :2.000                      Mode  :character
##                  Mean   :2.353
```

```
##           3rd Qu.:3.000
##           Max.    :4.000
## age12_plus_population age5_plus_population persons_fully_vaccinated
## Min.    :31651      Min.    : 36181      Min.    :13968
## 1st Qu.:37694      1st Qu.: 41612      1st Qu.:27447
## Median :43985      Median : 48573      Median :32322
## Mean   :46847      Mean   : 52012      Mean   :34364
## 3rd Qu.:53932      3rd Qu.: 59168      3rd Qu.:39176
## Max.    :88557      Max.    :101902     Max.    :71078
## persons_partially_vaccinated percent_of_population_fully_vaccinated
## Min.    : 1862      Min.    :0.3519
## 1st Qu.: 2853      1st Qu.:0.5891
## Median : 3532      Median :0.6649
## Mean   : 3917      Mean   :0.6630
## 3rd Qu.: 4524      3rd Qu.:0.7286
## Max.    :14941      Max.    :1.0000
## percent_of_population_partially_vaccinated
## Min.    :0.04658
## 1st Qu.:0.06103
## Median :0.06941
## Mean   :0.07532
## 3rd Qu.:0.08301
## Max.    :0.33035
## percent_of_population_with_1_plus_dose   redacted
## Min.    :0.4168      Length:411
## 1st Qu.:0.6677      Class :character
## Median :0.7377      Mode  :character
## Mean   :0.7371
## 3rd Qu.:0.8068
## Max.    :1.0000
```

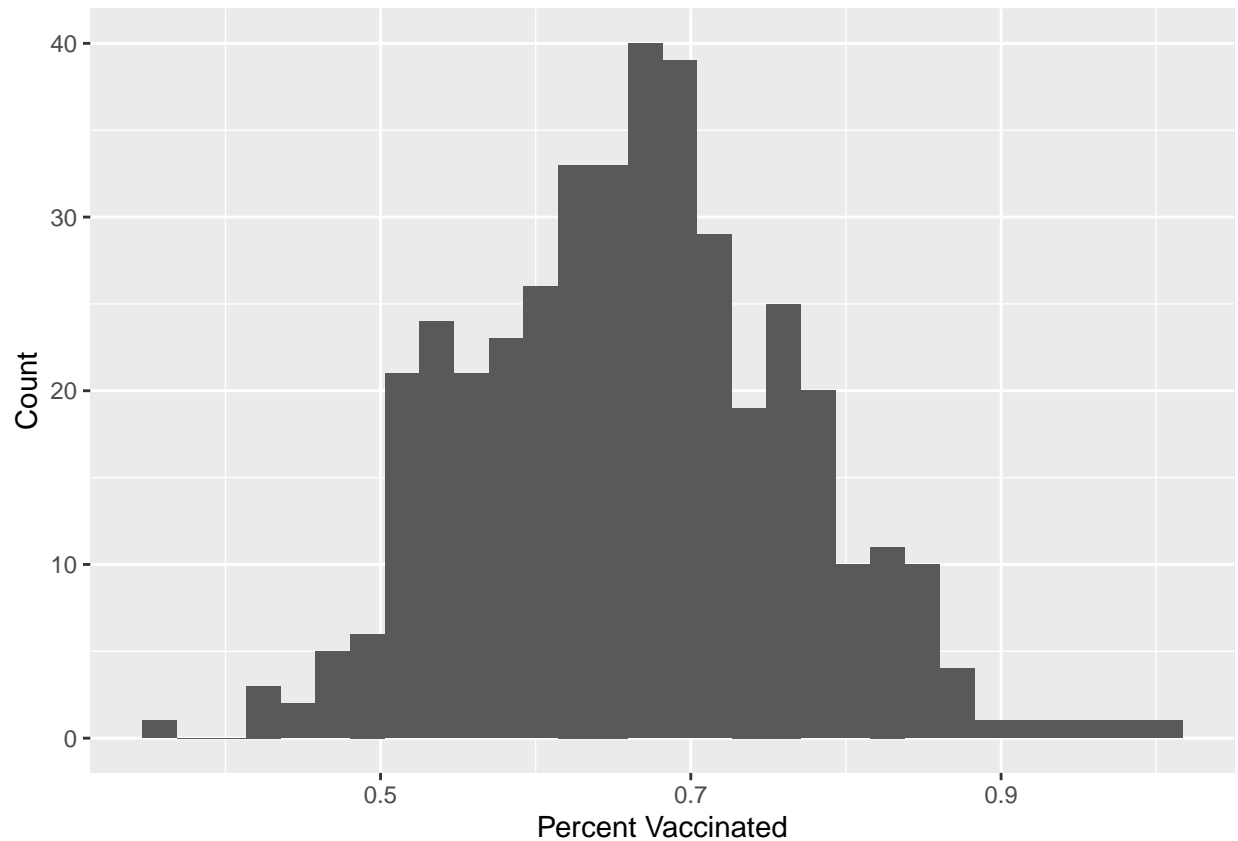
gasjfladfvkjerhgdwtkejhrk

Q18. Using ggplot generate a histogram of this data.

```
library(ggplot2)

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

```
## `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
```

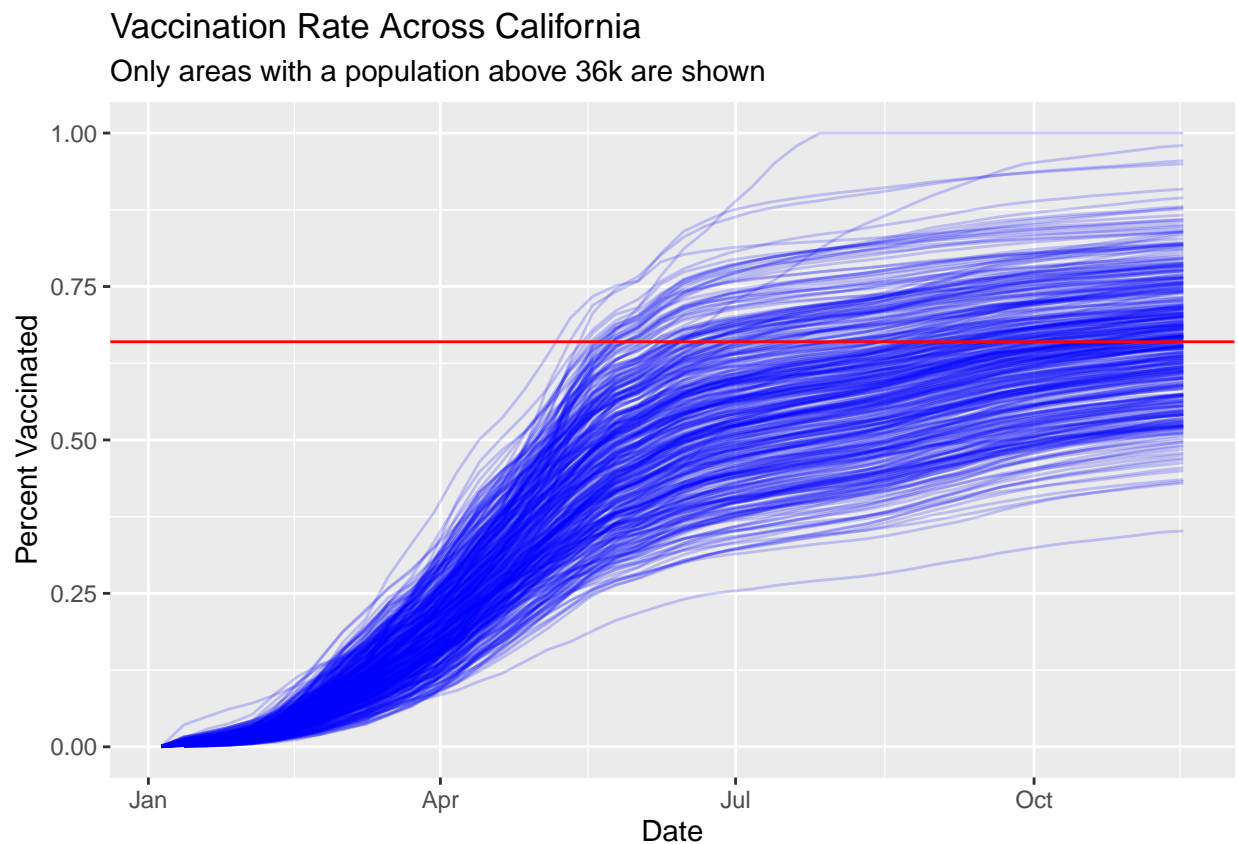
The zip code 92040 is below the average and the zip code 92109 is above the average.

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") +
  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, col="red")
```

## 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'm staying in San Diego for the break, but since I can imagine that a lot of people will be going home for the break, I think I'd feel safer if class wasn't in-person.