

class17 mini project

faisal

```
# 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	95446	Sonoma	Sonoma
2	2021-01-05	96014	Siskiyou	Siskiyou
3	2021-01-05	96087	Shasta	Shasta
4	2021-01-05	96008	Shasta	Shasta
5	2021-01-05	95410	Mendocino	Mendocino
6	2021-01-05	95527	Trinity	Trinity

	vaccine_equity_metric_quartile	vem_source
1	2	Healthy Places Index Score
2	2	CDPH-Derived ZCTA Score
3	2	CDPH-Derived ZCTA Score
4	NA	No VEM Assigned
5	3	CDPH-Derived ZCTA Score
6	2	CDPH-Derived ZCTA Score

	age12_plus_population	age5_plus_population	tot_population
1	4840.7	5057	5168
2	135.0	135	135
3	513.9	544	544
4	1125.3	1164	NA
5	926.3	988	997
6	476.6	485	499

	persons_fully_vaccinated	persons_partially_vaccinated
1	NA	NA
2	NA	NA
3	NA	NA
4	NA	NA
5	NA	NA

6	NA	NA
	percent_of_population_fully_vaccinated	
1	NA	
2	NA	
3	NA	
4	NA	
5	NA	
6	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
	bivalent_dose_recip_count	eligible_recipient_count
1	NA	0
2	NA	0
3	NA	2
4	NA	2
5	NA	0
6	NA	0

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

vax\$as_of_date

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

[6] "2021-01-05" "2021-01-05" "2021-01-05" "2021-01-05" "2021-01-05"

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
[99986] "2022-02-01" "2022-02-01" "2022-02-01" "2022-02-01" "2022-02-01"
[99991] "2022-02-01" "2022-02-01" "2022-02-01" "2022-02-01" "2022-02-01"
[99996] "2022-02-01" "2022-02-01" "2022-02-01" "2022-02-01"
[ reached getOption("max.print") -- omitted 99333 entries ]
```

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

vax\$persons_fully_vaccinated > Q2. What column details the Zip code tabulation area?

vax\$persons_fully_vaccinated

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_data[nrow(vax)]
```

NULL

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

Table 1: Data summary

Name	vax
Number of rows	199332
Number of columns	18
Column type frequency:	
character	5
numeric	13
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	113	0
local_health_jurisdiction	0	1	0	15	565	62	0
county	0	1	0	15	565	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	1817.38	0	192257.75	3658.50	5380.50	7635.0	
vaccine_equity_metric_0831tile	0	0.95	2.44	1.11	1	1.00	2.00	3.00	4.0	
age12_plus_population	0	1.00	18895.04	8993.87	0	1346.95	13685.13	1756.18	8556.7	
age5_plus_population	0	1.00	20875.24	1105.97	0	1460.50	15364.06	1877.00	1902.0	
tot_population	9718	0.95	23372.77	2628.51	2	2126.00	18714.08	168.00	1165.0	
persons_fully_vaccinated	16525	0.92	13962.33	5054.09	1	930.00	8566.00	23302.08	7566.0	
persons_partially_vaccinated	16525	0.92	1701.64	2030.18	11	165.00	1196.00	2535.00	39913.0	
percent_of_population_fully_vaccinated	20825	0.90	0.57	0.25	0	0.42	0.60	0.74	1.0	
percent_of_population_partially_vaccinated	20825	0.90	0.08	0.09	0	0.05	0.06	0.08	1.0	
percent_of_population_1_plus_dose	21859	0.89	0.63	0.24	0	0.49	0.67	0.81	1.0	
booster_recip_count	72872	0.63	5837.31	7165.81	11	297.00	2748.00	9438.25	9553.0	
bivalent_dose_recip_count	158664	0.20	2924.93	3583.45	11	190.00	1418.00	4626.25	7458.0	
eligible_recipient_count	0	1.00	12801.84	4908.33	0	504.00	6338.00	21973.08	7234.0	

Q5. How many numeric columns are in this dataset?

13

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] 16525

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

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

```
[1] 0.08290189
```

```
library(lubridate)
```

Attaching package: 'lubridate'

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

date, intersect, setdiff, union

```
today()
```

```
[1] "2023-03-07"
```

We can now magically do math with dates

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

```
today() - ymd("2021-01-05")
```

Time difference of 791 days

```
today() - ymd("2002-06-12")
```

Time difference of 7573 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

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

```

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_~1 major_~2 post_~3 common_c~4 county state lat lng timez~5
  <chr>   <chr>      <chr>   <chr>      <blob> <chr>   <chr> <dbl> <dbl> <chr>
1 92037   Standard   La Jol~ La Jol~ <raw 20 B> San D~ CA    32.8 -117. Pacific
2 92109   Standard   San Di~ San Di~ <raw 21 B> San D~ CA    32.8 -117. Pacific
# ... with 14 more variables: 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>, and abbreviated variable names
#   1: zipcode_type, 2: major_city, 3: post_office_city, ...

#Focus on the San Diego area

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

[1] 12091

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] 12091

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

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

```
n_distinct(sd$zip_code_tabulation_area)
```

[1] 107

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

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

[1] 92154

Q13. What is the overall average “Percent of Population Fully Vaccinated” value for all San Diego “County” as of “2022-11-15”?

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

[1] "2023-02-28"


```
sd.today <- filter(sd, as_of_date == "2023-02-28")
```

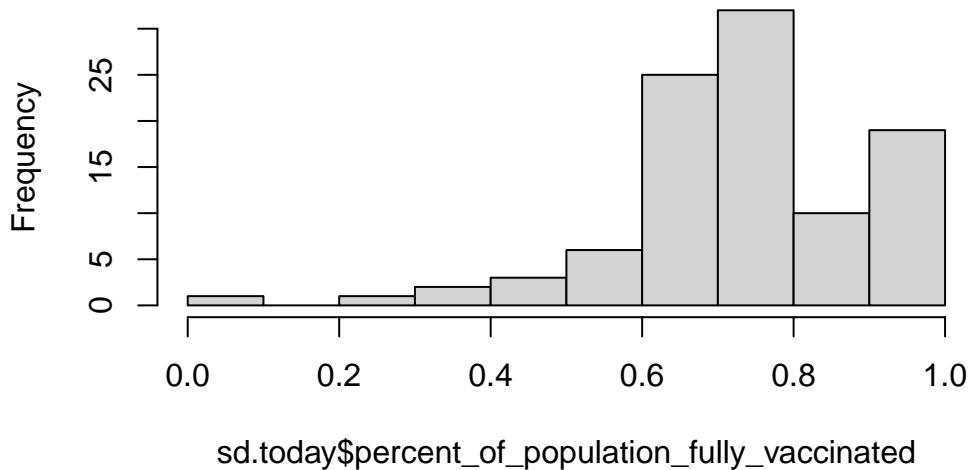
```
mean(sd.today$percent_of_population_fully_vaccinated, na.rm=T)
```

```
[1] 0.7400878
```

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 “2022-11-15”?

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

Histogram of sd.today\$percent_of_population_fully_vaccinated

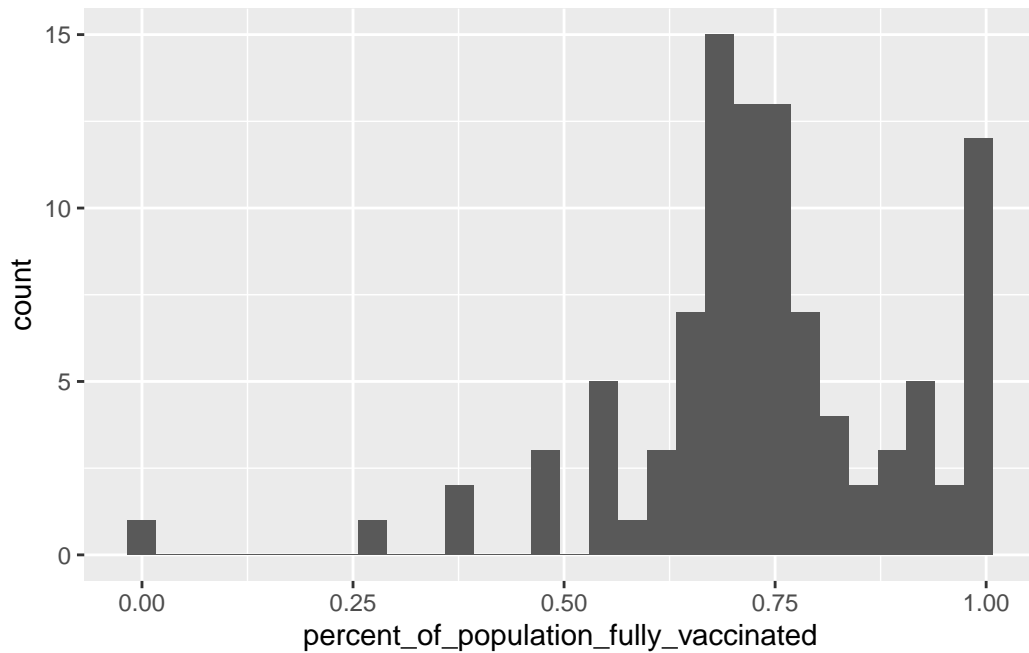


```
library(ggplot2)
```

```
ggplot(sd.today) +  
  aes(percent_of_population_fully_vaccinated) +  
  geom_histogram()
```

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

Warning: Removed 8 rows containing non-finite values (`stat_bin()`).



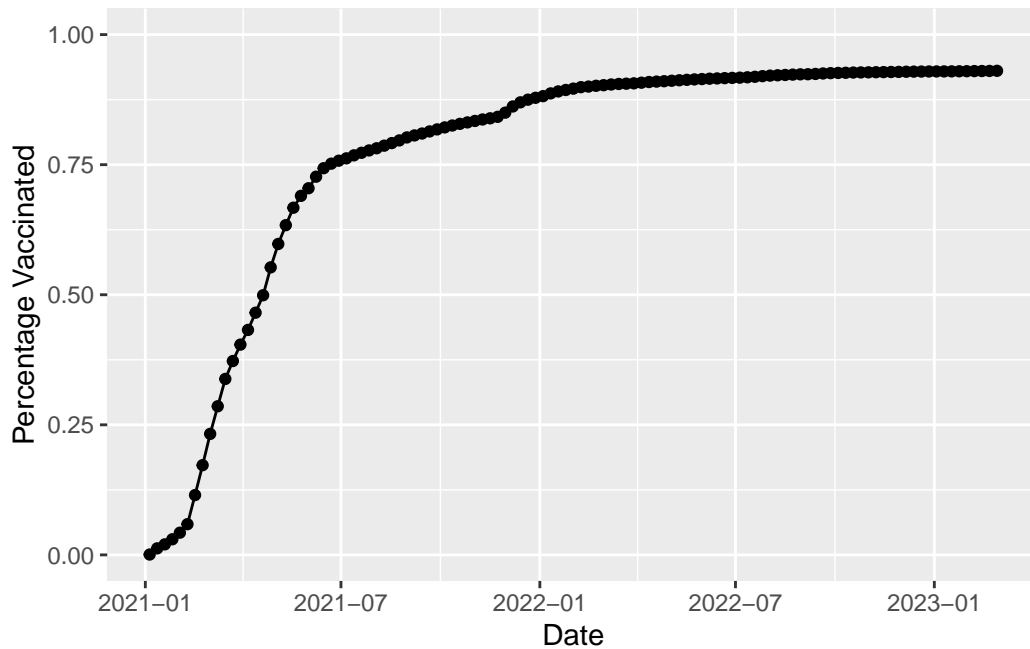
```
ucsd <- filter(sd, zip_code_tabulation_area=="92037")
ucsd[1,]$age5_plus_population
```

[1] 36144

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

```
Ucplot <- 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="Percentage Vaccinated")
```

Ucplot



##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 == "2023-02-28")

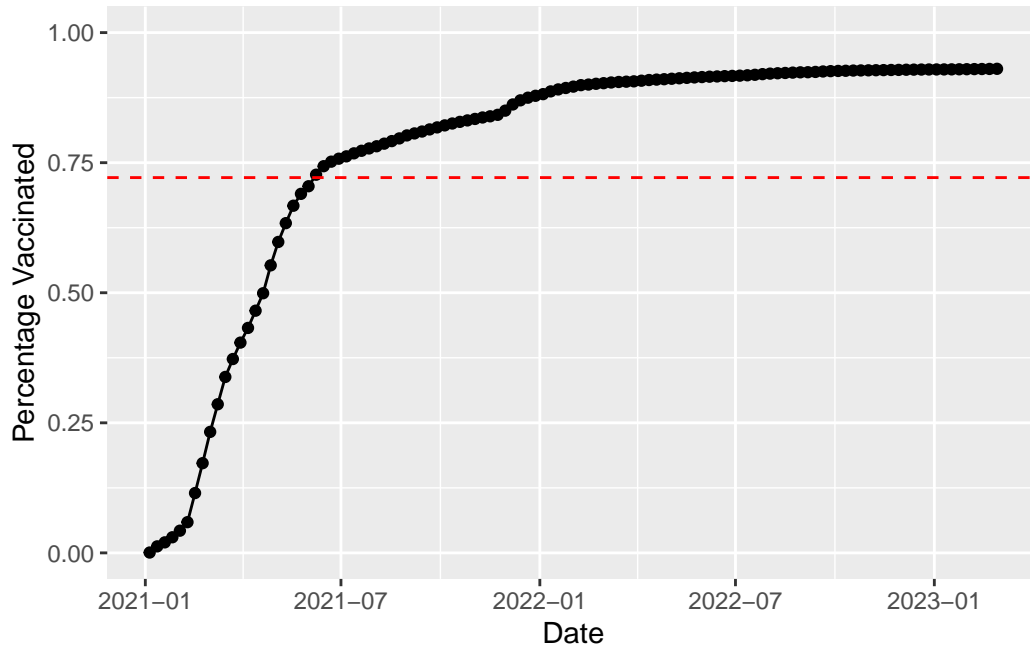
#head(vax.36)
```

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 “2022-11-15”. Add this as a straight horizontal line to your plot from above with the `geom_hline()` function?

```
ave <- mean(vax.36$percent_of_population_fully_vaccinated)
ave
```

```
[1] 0.7213331
```

```
Ucplot + geom_hline(yintercept=ave, col = "red", linetype =2)
```



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 “2022-11-15”?

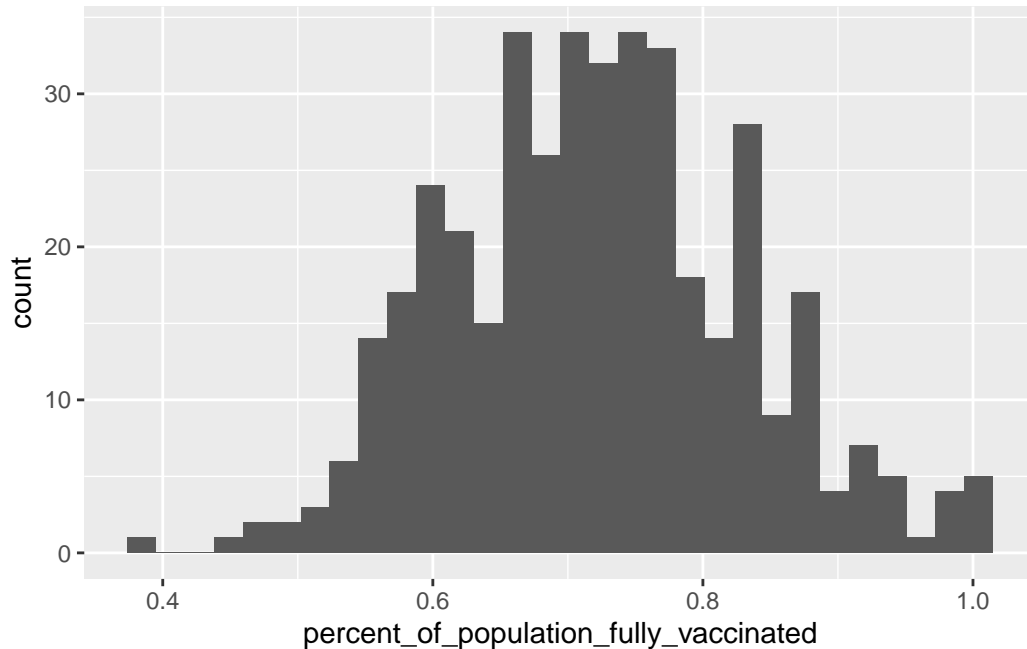
```
summary(vax.36$percent_of_population_fully_vaccinated)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.3804	0.6457	0.7181	0.7213	0.7907	1.0000

Q18. Using ggplot generate a histogram of this data.

```
ggplot(vax.36) +
  aes(percent_of_population_fully_vaccinated)+
  geom_histogram()
```

``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 == "2022-11-15") %>%
  filter(zip_code_tabulation_area=="92040") %>%
  select(percent_of_population_fully_vaccinated)
```

```
percent_of_population_fully_vaccinated
1                                0.548849
```

```
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= "Percentage Vaccinated",
       title= "Vaccination Rate Across America",
       subtitle="only Areas with a population above 36k are Shown") +
```

```
geom_hline(yintercept = mean(vax.36$percent_of_population_fully_vaccinated), linetype=2)
```

Warning: Removed 183 rows containing missing values (`geom_line()`).

