

## Milestone #3

Virginia Chan, Patrick Traynor, Antoinette Stetzenmeyer

11/4/2021

```
library(readr)
cov_vax_admin <- read_csv("cov_vax_admin.csv")

##
## -- Column specification -----
## cols(
##   X1 = col_double(),
##   as_of_date = col_character(),
##   zip_code_tabulation_area = col_double(),
##   local_health_jurisdiction = col_character(),
##   county = col_character(),
##   vaccine_equity_metric_quartile = col_double(),
##   vem_source = col_character(),
##   age12_plus_population = col_double(),
##   persons_fully_vaccinated = col_double(),
##   persons_partially_vaccinated = col_double(),
##   redacted = col_character()
## )
cov_vax_admin

## # A tibble: 65,268 x 11
##       X1 as_of_date zip_code_tabulat~ local_health_jur~ county vaccine_equity_m~
##   <dbl> <chr>          <dbl> <chr>          <chr>          <dbl>
## 1     1 1/5/2021          92703 ORANGE      ORANGE          1
## 2     2 1/5/2021          92285 SAN BERNARDINO SAN B~          1
## 3     3 1/5/2021          92284 SAN BERNARDINO SAN B~          1
## 4     4 1/5/2021          92275 IMPERIAL      IMPER~          1
## 5     5 1/5/2021          92532 RIVERSIDE      RIVER~          3
## 6     6 1/5/2021          92376 SAN BERNARDINO SAN B~          1
## 7     7 1/5/2021          92345 SAN BERNARDINO SAN B~          1
## 8     8 1/5/2021          91343 LOS ANGELES    LOS A~          2
## 9     9 1/5/2021          91910 SAN DIEGO      SAN D~          2
## 10    10 1/5/2021          91773 LOS ANGELES    LOS A~          3
## # ... with 65,258 more rows, and 5 more variables: vem_source <chr>,
## #   age12_plus_population <dbl>, persons_fully_vaccinated <dbl>,
## #   persons_partially_vaccinated <dbl>, redacted <chr>
ca_county_demographics <- read_csv("ca_county_demographics.csv")

## Warning: Missing column names filled in: 'X1' [1]

##
## -- Column specification -----
## cols(
```

```
## .default = col_double(),
## name = col_character()
## )
## i Use 'spec()' for the full column specifications.
```

```
ca_county_demographics
```

```
## # A tibble: 58 x 23
##       X1 name pop2012 pop12_sqmi white black ameri_es asian hawn_pi hispanic
##       <dbl> <chr>   <dbl>   <dbl> <dbl> <dbl>   <dbl> <dbl>   <dbl>   <dbl>
## 1      1 Kern    851089    104.  5.00e5 48921   12676 3.48e4   1252  413033
## 2      2 Kings  155039    111.  8.30e4 11014    2562 5.62e3    271  77866
## 3      3 Lake   65253     49.1 5.20e4  1232    2049 7.24e2    108  11088
## 4      4 Lass~   35039      7.42 2.55e4  2834    1234 3.56e2    165   6117
## 5      5 Los ~ 9904341  2423.  4.94e6 856874  72828 1.35e6  26094 4687889
## 6      6 Made~  153025     71.1 9.45e4  5629    4136 2.80e3    162  80992
## 7      7 Marin  255509     486.  2.02e5  6987    1523 1.38e4    509  39069
## 8      8 Mari~  18455     12.6 1.61e4   138     527 2.04e2     26  1676
## 9      9 Mend~  88094     25.1 6.72e4   622    4277 1.45e3    119  19505
## 10    10 Merc~ 256841     130.  1.48e5  9926    3473 1.88e4    583 140485
## # ... with 48 more rows, and 13 more variables: other <dbl>, mult_race <dbl>,
## #   males <dbl>, females <dbl>, med_age <dbl>, households <dbl>,
## #   families <dbl>, hse_units <dbl>, ave_fam_sz <dbl>, vacant <dbl>,
## #   owner_occ <dbl>, renter_occ <dbl>, county_fips <dbl>
```

```
unique(cov_vax_admin$as_of_date)
```

```
## [1] "1/5/2021" "1/12/2021" "1/19/2021" "1/26/2021" "2/2/2021" "2/9/2021"
## [7] "2/16/2021" "2/23/2021" "3/2/2021" "3/9/2021" "3/16/2021" "3/23/2021"
## [13] "3/30/2021" "4/6/2021" "4/13/2021" "4/20/2021" "4/27/2021" "5/4/2021"
## [19] "5/11/2021" "5/18/2021" "5/25/2021" "6/1/2021" "6/8/2021" "6/15/2021"
## [25] "6/22/2021" "6/29/2021" "7/6/2021" "7/13/2021" "7/20/2021" "7/27/2021"
## [31] "8/3/2021" "8/10/2021" "8/17/2021" "8/24/2021" "8/31/2021" "9/7/2021"
## [37] "9/14/2021"
```

#1 Subset rows or columns, as needed

*#Subset cov\_vax\_admin dataset by county, zipcode, and date*

```
total_age12andabove_california_county<- cov_vax_admin %>%  
  select(c(as_of_date, zip_code_tabulation_area, county, vaccine_equity_metric_quartile, age12_plus_pop  
  group_by(county, zip_code_tabulation_area, as_of_date) %>%  
  arrange(county)%>%  
summarize(total_age12andabove = sum(age12_plus_population,na.rm = TRUE), median_of_age_12_and_above = m
```

## 'summarise()' has grouped output by 'county', 'zip\_code\_tabulation\_area'. You can override using the

```
total_age12andabove_california <- cov_vax_admin %>%  
  group_by(county)%>%  
summarize(total_age12andabove = sum(age12_plus_population,na.rm = TRUE), median_of_age_12_and_above = m
```

#2 Create new variables needed for analysis (minimum 2) New variables should be created based on existing columns; for example Calculating a rate, Combining character strings Etc If no new values are needed for final tables/graphs, please create 2 new variables anyway

#3 Clean variables needed for analysis (minimum 2) Examples Recode invalid values Handle missing fields Recode categories Etc. If not needed for final analysis, please create at least 2 new variables anyway

#4 Data dictionary based on clean dataset (minimum 4 data elements), including: Variable name Data type  
Description

#5 One or more tables with descriptive statistics for 4 data elements

#6 PDF that is professionally prepared for presentation Each part of the milestone is clearly on one page  
(use



to push to a new page) Only the necessary information is outputted (you should suppress, for example, entire data frame outputs) Use of headers and sub headers to create an organized document