Milestone #3

Virginia Chan, Patrick Traynor, Antoinette Stetzenmeyer

11/8/2021

This is the code used to load the two data sets of interest: cox_vax_admin.cvs and ca_county_demographics.csv. The first 10 rows are provided to provide a sample of the variables and values

```
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()
## )
```

```
head(cov_vax_admin)
```

```
## # A tibble: 6 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 BE~                  1
## 3     3 1/5/2021               92284 SAN BERNARDINO    SAN BE~                  1
## 4     4 1/5/2021               92275 IMPERIAL          IMPERI~                  1
## 5     5 1/5/2021               92532 RIVERSIDE         RIVERS~                  3
## 6     6 1/5/2021               92376 SAN BERNARDINO    SAN BE~                  1
## # ... with 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.
```

```
head(ca_county_demographics)
```

```
## # A tibble: 6 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 Lassen   35039       7.42 2.55e4   2834     1234 3.56e2     165     6117
## 5     5 Los A~ 9904341    2423.   4.94e6 856874    72828 1.35e6   26094  4687889
## 6     6 Madera  153025      71.1  9.45e4   5629     4136 2.80e3     162    80992
## # ... with 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>
```

This is to identify all of the unique dates in the cov_vax_admin.csv data frame. It shows cumulative totals, that is, prior to, the date provided.

```
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"
```

We can see that the first date is January 5, 2021 and the final date is September 14, 2021.

First Task: Subset rows or columns, as needed

```
total_age12andabove<- cov_vax_admin %>%
  filter(as_of_date == "9/14/2021") %>%
  group_by(county) %>%
  arrange(county)%>%
  drop_na(county) %>%
summarize(total_age12andabove = sum(age12_plus_population,na.rm = TRUE),
          persons_fully_vaccinated = sum(persons_fully_vaccinated,na.rm = TRUE))
```

We are merging the 2 datasets. For the demographics dataset, we are only keep the columns for county and population as of 2012. Then, we merged this dataset with COVID vaccinations dataset.

```
total_pop <- ca_county_demographics %>%
  select(name, pop2012) %>%
  mutate(name=str_to_upper(name))

merged_dataset_pop_vax <- inner_join(total_age12andabove, total_pop, by=c("county"="name"))
```

Second Task: Create new variables needed for analysis (minimum 2)

We created 2 new variables: elg_vax_rate and ovrl_vax_rate.

```
merged_dataset_pop_vax_new_variables <- merged_dataset_pop_vax %>%
  mutate(elg_vax_rate=(persons_fully_vaccinated/total_age12andabove)) %>%
  mutate(ovrl_vax_rate=(persons_fully_vaccinated/pop2012))
```

Third Task Clean variables needed for analysis (minimum 2)

Here, we update the county column to be title case and rounded the 2 new variables to be displayed as percentages with 2 decimal place.

```
cleaned_merged_dataset_pop_vax_new_variables <-
  merged_dataset_pop_vax_new_variables %>%
  mutate(county=str_to_title(county)) %>%
  mutate(elg_vax_rate=round((elg_vax_rate) *100,2),
         ovrl_vax_rate=round((ovrl_vax_rate) *100,2))
```

Fourth Task: Data dictionary based on clean dataset (minimum 4 data elements), including:

Variable name Data type Description

Below is a data dictionary descibing all the variables used for this analysis.

```
kable(data_dict_age12_pl, format = "pipe", booktabs = T, caption = "Data
      Dictionary for COVID-19 Vaccination Rates for California from January to
      Mid September")
```

Table 1: Data Dictionary for COVID-19 Vaccination Rates for California from January to Mid September

Variable_Name	Data_Type	Description
county	Character	County Name
total_age12andabove	Double	Eligble vaccinated population (12 years & above)
persons_fully_vaccinated	Double	Number of individuals that are fully vaccinated
pop2012	Double	Latest census taken in year 2012
elg_vax_rate	Double	Percent of eligible population that is fully vaccinated
ovrl_vax_rate	Double	Percent of overall population that is fully vaccinated

Fifth Task: One or more tables with descriptive statistics for 4 data elements

We calculated the mean, maximum, minimum, and standard deviation for the 2 new variables of persons fully vaccinated and person who were eligible to receive a vaccine.

```
data_table_of_descriptives <- data.frame(
  Variable_Names = c("min","max","mean","standard deviation"),
  Persons_fully_vaccinated =
    c(min(cleaned_merged_dataset_pop_vax_new_variables$persons_fully_vaccinated),
  max(cleaned_merged_dataset_pop_vax_new_variables$persons_fully_vaccinated),
  mean(cleaned_merged_dataset_pop_vax_new_variables$persons_fully_vaccinated),
  sd(cleaned_merged_dataset_pop_vax_new_variables$persons_fully_vaccinated)),
  Pop_2012 = c(min(cleaned_merged_dataset_pop_vax_new_variables$pop2012),
  max(cleaned_merged_dataset_pop_vax_new_variables$pop2012),
  mean(cleaned_merged_dataset_pop_vax_new_variables$pop2012),
  sd(cleaned_merged_dataset_pop_vax_new_variables$pop2012)),
  Eligible_vax_rate = c(min(cleaned_merged_dataset_pop_vax_new_variables$elg_vax_rate),
  max(cleaned_merged_dataset_pop_vax_new_variables$elg_vax_rate),
  mean(cleaned_merged_dataset_pop_vax_new_variables$elg_vax_rate),
  sd(cleaned_merged_dataset_pop_vax_new_variables$elg_vax_rate)))

kable(data_table_of_descriptives, booktabs = T, format = "pipe", caption =
"This shows the minimum, maximum, mean, and standard deviation for the variables included")
```

Table 2: This shows the minimum, maximum, mean, and standard deviation for the variables included

Variable_Names	Persons_fully_vaccinated	Pop_2012	Eligible_vax_rate
min	416.0	1148.0	22.21000
max	5913908.0	9904341.0	86.26000
mean	387965.9	650128.9	58.26845
standard deviation	867644.4	1431319.1	14.19526