Project Milestone 2

Authors: Kathy Nazarian, Kimberly Michel

```
#packages needed
library(readr)
library(lubridate)
```

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

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

```
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
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats 1.0.0     v stringr 1.5.0
## v ggplot2 3.4.2     v tibble  3.2.1
## v purrr   1.0.2     v tidyr   1.3.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

Import Data Sets

```
ca_vax_rates_quarter <- read_csv("ca_vax_rates_quarter.csv")
```

```
## Rows: 13533 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): County Name, County Type, Demographic Category, Demographic Value
## dbl  (6): Estimated Population, Total Partial Vaccinated, Cumulative Fully V...
## lgl  (1): Suppress Data
## date (2): Dt Admin, Quarter
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
ca_vax <- ca_vax_rates_quarter

sim_flu_CA <- read_csv("sim_flu_CA.csv")
```

```
## Rows: 114912 Columns: 18
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (3): county, age_cat, sex
## dbl  (13): time_int, new_infections, new_recovered, count_susceptible, curre...
## date  (2): report_date, dt_diagnosis
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
CA_flu <- sim_flu_CA

sim_flu_LACounty <- read_csv("sim_flu_LACounty.csv")
```

```
## Rows: 2016 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): DT_DX, AGE_CATEGORY, SEX, RACE_ETH
## dbl (8): DX_NEW, RECOVERED_NEW, INFECTED_CURRENT, INFECTED_CUMULATIVE, RECOV...
## lgl (1): DT_REPORT
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
LACounty_flu <- sim_flu_LACounty
```

Clean Data using Snake Case

```r
names(ca_vax)[1] <- 'county_name'
names(ca_vax)[2] <- 'county_type'
names(ca_vax)[3] <- 'demographic_category'
names(ca_vax)[4] <- 'demographic_value'
names(ca_vax)[5] <- 'est_population'
names(ca_vax)[6] <- 'date_admin'
names(ca_vax)[7] <- 'total_partial_vax'
names(ca_vax)[8] <- 'cum_fully_vax'
names(ca_vax)[9] <- 'cum_atleast_one_dose'
names(ca_vax)[10] <- 'cum_unvax'
names(ca_vax)[11] <- 'suppress_data'
names(ca_vax)[12] <- 'cum_up_to_date_vax'
names(ca_vax)[13] <- 'quarter'

names(LACounty_flu)[1] <- 'date'
names(LACounty_flu)[2] <- 'age_category'
names(LACounty_flu)[3] <- 'sex'
names(LACounty_flu)[4] <- 'race_eth'
names(LACounty_flu)[5] <- 'dt_report'
names(LACounty_flu)[6] <- 'dx_new'
names(LACounty_flu)[7] <- 'recovered_new'
names(LACounty_flu)[8] <- 'infected_current'
names(LACounty_flu)[9] <- 'infected_cum'
names(LACounty_flu)[10] <- 'recovered_cum'
names(LACounty_flu)[11] <- 'susceptible'
names(LACounty_flu)[12] <- 'severe_new'
names(LACounty_flu)[13] <- 'severe_cum'
```

Data Cleaning for LACounty_flu ds

```
#dt_dx string is 'chr', must be changed to date

str(LACounty_flu)
```

```
## spc_tbl_ [2,016 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ date            : chr [1:2016] "03OCT2022" "09OCT2022" "19OCT2022" "26OCT2022" ...
##  $ age_category    : chr [1:2016] "0-17" "0-17" "0-17" "0-17" ...
##  $ sex             : chr [1:2016] "FEMALE" "FEMALE" "FEMALE" "FEMALE" ...
##  $ race_eth        : chr [1:2016] "White, Non-Hispanic" "White, Non-Hispanic" "White, Non-Hispanic"
##  $ dt_report       : logi [1:2016] NA NA NA NA NA NA ...
##  $ dx_new          : num [1:2016] 513 343 340 422 321 ...
##  $ recovered_new   : num [1:2016] 0 508 341 341 421 ...
##  $ infected_current: num [1:2016] 513 348 347 428 328 ...
##  $ infected_cum    : num [1:2016] 513 856 1196 1618 1939 ...
##  $ recovered_cum   : num [1:2016] 0 508 849 1190 1611 ...
##  $ susceptible     : num [1:2016] 215280 214937 214597 214175 213854 ...
##  $ severe_new      : num [1:2016] 0 0 0 1 1 0 1 0 1 1 ...
##  $ severe_cum      : num [1:2016] 0 0 0 1 2 2 3 3 4 5 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   DT_DX = col_character(),
##   ..   AGE_CATEGORY = col_character(),
##   ..   SEX = col_character(),
##   ..   RACE_ETH = col_character(),
##   ..   DT_REPORT = col_logical(),
##   ..   DX_NEW = col_double(),
##   ..   RECOVERED_NEW = col_double(),
##   ..   INFECTED_CURRENT = col_double(),
##   ..   INFECTED_CUMULATIVE = col_double(),
##   ..   RECOVERED_CUMULATIVE = col_double(),
##   ..   SUSCEPTIBLE = col_double(),
##   ..   SEVERE_NEW = col_double(),
##   ..   SEVERE_CUMULATIVE = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
#reformat date in LACounty_flu dataset from chr to date format
LACounty_flu <- LACounty_flu %>%
  mutate(date = as.Date(as.character(LACounty_flu$date), format = "%d%b%Y"))

#data cleaned by descending order by dt_dx
LACounty_flu <- LACounty_flu[order(LACounty_flu$date),]
```

Data Cleaning for CA_flu ds

```
# data cleaning for CA_flu ds
#data cleaned by descending order of diagnosis date
CA_flu <- CA_flu[order(CA_flu$dt_diagnosis),]
```

Data Cleaning for ca_vax ds

```
#data cleaning for ca_vax ds

#data reorganized by descending order of date vax was given
ca_vax <- ca_vax[order(ca_vax$date_admin),]
```

Structure of each df:

```r
str(CA_flu)
```

```
## tibble [114,912 x 18] (S3: tbl_df/tbl/data.frame)
##  $ county                 : chr [1:114912] "Butte County" "Colusa County" "San Bernardino County" "Ora
##  $ time_int               : num [1:114912] 202240 202240 202240 202240 202240 ...
##  $ new_infections         : num [1:114912] 0 0 21 40 97 1 3 1 3 0 ...
##  $ new_recovered          : num [1:114912] 0 0 0 0 0 0 0 0 0 0 ...
##  $ count_susceptible      : num [1:114912] 43 49 15495 18804 46376 ...
##  $ current_infected       : num [1:114912] 0 0 21 40 97 1 3 1 3 0 ...
##  $ cumulative_infected    : num [1:114912] 0 0 21 40 97 1 3 1 3 0 ...
##  $ cumulative_recovered   : num [1:114912] 0 0 0 0 0 0 0 0 0 0 ...
##  $ new_unrecovered        : num [1:114912] 0 0 0 0 0 0 0 0 0 0 ...
##  $ cumulative_unrecovered: num [1:114912] 0 0 0 0 0 0 0 0 0 0 ...
##  $ new_severe             : num [1:114912] 0 0 0 0 0 0 0 0 0 0 ...
##  $ cumulative_severe      : num [1:114912] 0 0 0 0 0 0 0 0 0 0 ...
##  $ age_cat                : chr [1:114912] "0-17" "18-49" "50-64" "18-49" ...
##  $ sex                    : chr [1:114912] "MALE" "FEMALE" "FEMALE" "FEMALE" ...
##  $ race_ethnicity         : num [1:114912] 5 4 4 6 7 1 4 7 6 5 ...
##  $ report_date            : Date[1:114912], format: "2022-10-08" "2022-10-08" ...
##  $ dt_diagnosis           : Date[1:114912], format: "2022-09-24" "2022-09-24" ...
##  $ pop                    : num [1:114912] 43 49 15495 18804 46376 ...
```

```r
str(LACounty_flu)
```

```
## tibble [2,016 x 13] (S3: tbl_df/tbl/data.frame)
##  $ date            : Date[1:2016], format: "2022-09-29" "2022-09-29" ...
##  $ age_category    : chr [1:2016] "0-17" "65+" "0-17" "18-49" ...
##  $ sex             : chr [1:2016] "FEMALE" "MALE" "FEMALE" "MALE" ...
##  $ race_eth        : chr [1:2016] "Asian, Non-Hispanic" "White, Non-Hispanic" "Multiracial (two or mo
##  $ dt_report       : logi [1:2016] NA NA NA NA NA NA ...
##  $ dx_new          : num [1:2016] 271 671 89 5 270 ...
##  $ recovered_new   : num [1:2016] 0 0 0 0 0 0 0 0 0 0 ...
##  $ infected_current: num [1:2016] 271 671 89 5 270 ...
##  $ infected_cum    : num [1:2016] 271 671 89 5 270 ...
##  $ recovered_cum   : num [1:2016] 0 0 0 0 0 0 0 0 0 0 ...
##  $ susceptible     : num [1:2016] 112939 311632 37348 3828 131576 ...
##  $ severe_new      : num [1:2016] 0 0 0 0 0 0 0 0 0 0 ...
##  $ severe_cum      : num [1:2016] 0 0 0 0 0 0 0 0 0 0 ...
```

```r
str(ca_vax)
```

```
## tibble [13,533 x 13] (S3: tbl_df/tbl/data.frame)
##  $ county_name         : chr [1:13533] "Santa Barbara" "Humboldt" "Mendocino" "Kings" ...
##  $ county_type         : chr [1:13533] "MIXED" "MIXED" "MIXED" "MIXED" ...
##  $ demographic_category: chr [1:13533] "Gender" "Gender" "Gender" "Gender" ...
##  $ demographic_value   : chr [1:13533] "Male" "Female" "Male" "Female" ...
##  $ est_population      : num [1:13533] 229031 66782 44179 70157 20715 ...
##  $ date_admin          : Date[1:13533], format: "2020-08-04" "2020-08-05" ...
##  $ total_partial_vax   : num [1:13533] 2 1 1 1 1 1 1 1 3 3 ...
##  $ cum_fully_vax       : num [1:13533] 0 0 0 0 0 1 0 0 1 0 ...
##  $ cum_atleast_one_dose: num [1:13533] 2 1 1 1 1 2 1 1 4 3 ...
##  $ cum_unvax           : num [1:13533] 229029 66781 44178 70156 20714 ...
##  $ suppress_data       : logi [1:13533] FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ cum_up_to_date_vax  : num [1:13533] 0 0 0 0 0 0 0 0 0 0 ...
##  $ quarter             : Date[1:13533], format: "2020-07-01" "2020-07-01" ...
```

Data Type Descriptions

```
summary(ca_vax_rates_quarter$`Estimated Population`)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
##        0     3087    18116   271526   110334 20036369     2061
```

```
summary(ca_vax_rates_quarter$`Estimated Population`)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
##        0     3087    18116   271526   110334 20036369     2061
```

```
summary(ca_vax_rates_quarter$`Cumulative Fully Vaccinated`)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
##        0      188     3674   133568    33426 15143943      923
```

```
summary(ca_vax_rates_quarter$`Cumulative Unvaccinated`)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
##        0     1794     9101   120952    44255 20035473     2893
```

```
summary(sim_flu_CA$new_infections)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0     0.0     2.0   112.6    28.0 12672.0
```

```
summary(sim_flu_CA$new_severe)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.000   0.000   0.000   2.511   0.000 562.000
```

Description of dataset:

Datasets appear like they were collected by a health department within CA or Los Angeles. The data that was collected includes cases of flu as well as vaccine uptake data. Timeframe of each dataset ranges between 2020-2023 within different counties across California.

Question/Statement: Reporting whether there's any correlation between flu vaccination rates and covid vaccination rates.

The dataset provides vaccination rates within many counties in california as well as the amount of cases. This will allow us to conduct our analysis in order to determine whether there's any correlation between flu and covid vaccination rates by running appropriate statistical tests.

5+ data types:

-Population sizes This information is essential as it'll allow us to conduct our calculations in order to determine rates.

-Cumulative up to date vaccinated/unvaccinated (exposure) Will be using this data to calculate vaccination rates in order to compare them to flu rates among exposed and unexposed

-New severe Will be using this data (incidence rate) to compare to exposed and unexposed populations (vaccinated/unvaccinated)

-County Could be a confounding variable. Could assess whether some counties have higher vaccination/flu rates which would be considered a disparity.

-Age Could be a counting variable as well. We could use direct standardization in order to see the crude rates among different age groups in order to assess whether certain age groups have higher rates than others.

Data cleaning:

Completed: -renamed all dataset column names by snake case -Changed LACounty flu "dt_dx" column from chr string to date, then arranges in chronological order + removed extra columns + moved "date" column to first

Need to be cleaned: -Demographic value needs to be reformatted from chr to num (ca_vax) -age needs to be reformatted from chr to num (LACounty_flu) -Age_cat needs to be reformatted from chr to number (ca_flu) -Ethnicity needs to be reformatted from num to chr (ca_flu)