# PHW251 Fall 2022 Project Milestone § 3

Aubrey Robinson, Elizabeth Guzman, Tin Ho (Group Z)

Due: 2022-11-07

## Project and Milestone overview

This is Milestone #3 of the semester-long project.

Project and Milestone overview pasted to group's gdoc as checklist

## R setup and load libraries

(this code block omitted for brevity)

### Loading data

(Code block omitted for brevity, mostly code from Milestone 2)

# Subset rows or columns, as needed

## Subset mortality data by Chronic diseases

For this project, we are asked to focus on the chronic diseases. The mortality data set include non chronic diseases.

After discussion in ed.com, https://edstem.org/us/courses/25507/discussion/2060979 , the chronic disease list is defined as below, which seems to agree with CDC definition as well

"Chronic lower respiratory diseases", "Diabetes mellitus", "Diseases of heart", "Essential hypertension and hypertensive renal disease", "Chronic liver disease and cirrhosis", "Alzheimer's disease", "Malignant neoplasms", "Nephritis, nephrotic syndrome and nephrosis", "Cerebrovascular diseases", "Parkinson's disease"

```
chronic_desc = c(
  "Chronic lower respiratory diseases",
  "Diabetes mellitus",
  "Diseases of heart",
  "Essential hypertension and hypertensive renal disease",
  "Chronic liver disease and cirrhosis",
  "Alzheimer's disease",
  "Malignant neoplasms",
  "Nephritis, nephrotic syndrome and nephrosis",
  "Cerebrovascular diseases",
  "Parkinson's disease"
  )


chronic_mortality_data = mortality_data %>%
  filter( Cause_Desc %in% chronic_desc )

num_rows_befor_filter = mortality_data %>% tally
num_rows_after_filter = chronic_mortality_data %>% tally

num_rows_filtered = num_rows_befor_filter - num_rows_after_filter
num_rows_filtered
```

```
##       n
## 1 58464
```

And we find that we remove 58,000+ rows of data that we don't need.

## Subsetting Funding and Demographics Data

We also need to review the most recent HCAI funding for projects in closure in each county. Below, is the funding_data subsetted for projects closed in August 2022

Other subsetting maybe needed, and we will develop them as we make progress on this project.

```
funding_data_full = funding_data
funding_data <- funding_data_full %>%
  group_by(County, Numeric_Cost, `OSHPD Project Status`) %>%
  filter(`OSHPD Project Status` == "In Closure") %>%
```

```
  filter(`Data Generation Date` == as_date("2022-08-11") )

demographics_data_subset <- demographics_data %>%
  select(name, pop2012, pop12_sqmi, med_age, owner_occ, renter_occ)
```

# Create New Variables

- 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

## Renters vs Homeowners

The following code block creates a new variable rent_own_ratio

```
## Renters vs Homeowners Ratio
demographics_data_subset = demographics_data_subset %>%
  mutate( rent_own_ratio = renter_occ / owner_occ )
```

## Rural Areas

For the second new variable, we are categorizing one of our existing variables as either "rural" or "not rural". We are using the National Rural Development Partnership's definition which counts an area with less than 20 people per square mile rural.

```
demographics_data_subset <-
  mutate(demographics_data_subset,
         rural_class=if_else( pop12_sqmi < 20, "rural", "not rural", missing=NULL))
```

## Clean variables needed for analysis (minimum 2)

In the funding_data table, County is coded like "19 - Los Angeles", "01 - Alameda". We need county name by itself for later table join process, thus we need to "clean out" the id portion in this data field. We store this cleaned data in a new column called County.

We note that the county code is prefixed with 0 when they are single digits, thus it is always the first 5 chars that need to be stripped out.

```
funding_data = funding_data %>%
  mutate( county_name = str_sub( County, 6 )) %>%
  subset(select = -c(County)) %>%
  rename(County = county_name)
```

Various columns have number formats that are unsightly, using round() to create easier to read numbers.

```
funding_data$Numeric_Cost = round( funding_data$Numeric_Cost )

demographics_data_subset$rent_own_ratio = round(
  demographics_data_subset$rent_own_ratio, 4   )
```

### Joining Table

Now we join the tables democraphics_data_subset with chronic_mortality_data by their county names to create a new table called deomographics_chronic

```
demographics_chronic <- left_join(demographics_data_subset,
                                  chronic_mortality_data,
                                  by = c("name" = "County") ) %>%
  rename( County=name )
```

## Data dictionary based on clean dataset (minimum 4 data elements)

- We find all 3 tables have county names in it, which can be used as key for joining these tables. However, in 1 case we need to use a cleaned version of this column. Overall, county name is a character data field contained in :
  - demographics_data Name
  - funding_data County
  - mortality_data County and this is the variable used as key to join tables.

### Other variables of interest

name = character, Name of County

pop12_sqmi= numeric, Number of people per square mile in county

med_age = numeric, Median age of people in county

rent_own_ratio= numeric, The ratio of people who rent over those who own a home

Cause_Desc= character, Names of Chronic Diseases people in counties suffer from

Table 1: Rural Counties of California and their renter to homeownership ratio

| County | Class | Rent vs Own Ratio |
|---|---|---|
| Mono | rural | 0.7869 |
| Colusa | rural | 0.6341 |
| Inyo | rural | 0.5718 |
| Siskiyou | rural | 0.5445 |
| Lassen | rural | 0.5263 |
| Mariposa | rural | 0.4718 |
| Modoc | rural | 0.4587 |
| Plumas | rural | 0.4398 |
| Trinity | rural | 0.4199 |
| Alpine | rural | 0.3922 |
| Sierra | rural | 0.3915 |

Table 2: Number of People Suffering from Each Chronic Illness Type by County, 2014-2020

| County | ALZ | CAN | CLD | DIA | HTD | HYP | LIV | NEP | PAR | STK | Sum | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Siskiyou | 176 | 750 | 282 | 108 | 751 | 0 | 0 | 0 | 0 | 153 | 2,220 | 317 |
| Inyo | 0 | 231 | 93 | 0 | 272 | 11 | 0 | 0 | 0 | 0 | 607 | 87 |
| Plumas | 0 | 245 | 69 | 0 | 279 | 0 | 0 | 0 | 0 | 0 | 593 | 85 |
| Mariposa | 0 | 226 | 0 | 0 | 305 | 0 | 0 | 0 | 0 | 0 | 531 | 76 |
| Lassen | 0 | 198 | 0 | 0 | 296 | 0 | 0 | 0 | 0 | 0 | 494 | 71 |
| Colusa | 38 | 163 | 0 | 0 | 193 | 0 | 0 | 0 | 0 | 0 | 394 | 56 |
| Trinity | 0 | 135 | 11 | 0 | 181 | 0 | 0 | 0 | 0 | 0 | 327 | 47 |
| Modoc | 0 | 117 | 39 | 0 | 133 | 0 | 0 | 0 | 0 | 0 | 289 | 41 |
| Mono | 0 | 0 | 0 | 0 | 74 | 0 | 0 | 0 | 0 | 0 | 74 | 11 |
| Alpine | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Sierra | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

## One or more tables with descriptive statistics for 4 data element

First, we create a table for rural counties and their home rent:own ratio calculated previously, sorted by decreasing home ownership ratio.

Second, we produce a table for the total case count in 2014-2020 for the rural counties.

The “Sum” column is the total cases over all the chronic diseases for each county, and the “Mean” is the sum divided over the 7 years span. Together they provides rough yard stick of chronic cases each county oversees.

Note that there are many zeros in this table, they may not truly represent no cases, but more likely because of missing data.

Table 3: Disease Code Legend

| Cause | Cause Description |
|---|---|
| ALZ | Alzheimer's disease |
| CAN | Malignant neoplasms |
| CLD | Chronic lower respiratory diseases |
| DIA | Diabetes mellitus |
| HTD | Diseases of heart |
| HYP | Essential hypertension and hypertensive renal disease |
| LIV | Chronic liver disease and cirrhosis |
| NEP | Nephritis, nephrotic syndrome and nephrosis |
| PAR | Parkinson's disease |
| STK | Cerebrovascular diseases |