

# PHW251 Fall 2022 Project Milestone #2

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

Due: 2022-10-03

## Task for Milestone 2

Info from bCourse re-listed below as separate sections.

## Loading R libraries

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.1
```

```
## -- Attaching packages ----- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.7      v dplyr  1.0.9
## v tidyr   1.2.0      v stringr 1.4.0
## v readr   2.1.2      v forcats 0.5.2
```

```
## Warning: package 'ggplot2' was built under R version 4.2.1
```

```
## Warning: package 'tidyr' was built under R version 4.2.1
```

```
## Warning: package 'readr' was built under R version 4.2.1
```

```
## Warning: package 'forcats' was built under R version 4.2.1
```

```
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
library(dplyr)
```

## Description of dataset

- What is the data source? (1-2 sentences on where the data is coming from, dates included, etc.)

We have 3 different data sets that we will eventually be merging into a single one. They are county, mortality, and HCAI funding data. The county demographic data is from the 2012 California census. The mortality data was sourced from the CA Dept of Public Health's Death Profiles by County from 2014-2020. The HCAI funding data set comes from the CA.gov website which is the California open data portal, which includes data from 2013-2022.

- How does the dataset relate to the group problem statement and question?

The demographic dataset will help us identify counties that meet OHE's criteria of low population per square mile, high median age, and a high proportion of renters vs. homeowners. The mortality data will help the team identify counties with high occurrences of death caused by chronic health conditions. The HCAI funding dataset will help us identify which projects are in closure and which counties have been receiving a low amount of funding to help us determine which counties should receive more funding.

## Import statement

- NOTE: Please use datasets available in the PHW251 Project Data github repo (Links to an external site.) (this is important to make sure everyone is using the same datasets)
- Use appropriate import function and package based on the type of file
- Utilize function arguments to control relevant components (i.e. change column types, column names, missing values, etc.)
- Document the import process

## Loading data

We will use dplyr's read\_csv to import the data, since they are all presented as csv files.

```
# - Utilize function arguments to control relevant components
#   (i.e. change column types, column names, missing values, etc.)

demographics_path = 'data/ca_county_demographics.csv'
demographics_data = read_csv( demographics_path,
                              na = c("", "NA", "-"),
                              show_col_types=F )

## New names:
## * ' ' -> '...1'

# the first column contains id number, but it is unnamed, so renaming it
# rest of the columns have reasonable names, so left them as is.
demographics_data = rename( demographics_data, id="...1")

# View(demographics_data)

# Loading Mortality Data

mortality_path <- 'data/ca_county_mortality.csv'
mortality_data_raw <- read_csv(mortality_path,
                              na= c("", "NA", "-"),
                              show_col_types=F)

mortality_data <- mortality_data_raw %>% mutate_all(~replace( ., is.na(.), 0))

# View(mortality_data)
# There are NA values in mortality_data, so we need to replace NA w/ 0

# Loading HCAI funding Data

funding_path = 'data/hcai_healthcare_construction.csv'
funding_data <- read_csv( funding_path,
                          na= c("", "NA", "-"),
                          show_col_types=F )

# finding where in the data frame there is an 'na'
```

```

# https://www.geeksforgeeks.org/find-columns-and-rows-with-na-in-r-dataframe/
funding_data_no_CtyColl = select(funding_data, -c("Collection of Counties"))
which(is.na(funding_data_no_CtyColl), arr.ind=T)

##      row col

# and we find that only the column "Collection of Counties" has 'na'
# we will leave this for now since it may just be a colloquial reference
# unimportant for our data analysis.
# no replacement for na with 0 will be done on this data frame.

# the Costs column has human data, eg $50,890,315.00
# and we need to strip the dollar sign, the commas,
# and convert them to numbers.
# ref: https://stackoverflow.com/questions/31944103/convert-currency-with-commas-into-numeric
# we create a new column for this called "Numeric_Cost",
# but could have potentially done an in-place replacement
funding_data = funding_data %>%
  mutate(Numeric_Cost = as.numeric(
    gsub( '$,', '', funding_data[["Total Costs of OSHPD Projects"]] )
  ))

```

## Identify data types for 5+ data elements/columns/variables

- Identify 5+ data elements required for your specified scenario. If <5 elements are required to complete the analysis, please choose additional variables of interest in the data set to explore in this milestone.
- Utilize functions or resources in RStudio to determine the types of each data element (i.e. character, numeric, factor)
- Identify the desired type/format for each variable—will you need to convert any columns to numeric or another type?

```
str( demographics_data )
```

```
## spec_tbl_df [58 x 23] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
## $ id          : num [1:58] 1 2 3 4 5 6 7 8 9 10 ...
## $ name        : chr [1:58] "Kern" "Kings" "Lake" "Lassen" ...
## $ pop2012     : num [1:58] 851089 155039 65253 35039 9904341 ...
## $ pop12_sqmi  : num [1:58] 104.28 111.43 49.08 7.42 2423.26 ...
## $ white       : num [1:58] 499766 83027 52033 25532 4936599 ...
## $ black       : num [1:58] 48921 11014 1232 2834 856874 ...
## $ ameri_es    : num [1:58] 12676 2562 2049 1234 72828 ...
## $ asian       : num [1:58] 34846 5620 724 356 1346865 ...
## $ hawn_pi     : num [1:58] 1252 271 108 165 26094 ...
## $ hispanic    : num [1:58] 413033 77866 11088 6117 4687889 ...
## $ other       : num [1:58] 204314 42996 5455 3562 2140632 ...
## $ mult_race   : num [1:58] 37856 7492 3064 1212 438713 ...
## $ males       : num [1:58] 433108 86344 32469 22416 4839654 ...
## $ females     : num [1:58] 406523 66638 32196 12479 4978951 ...
## $ med_age     : num [1:58] 30.7 31.1 45 37 34.8 33.1 44.5 49.2 41.6 29.6 ...
## $ households  : num [1:58] 254610 41233 26548 10058 3241204 ...
## $ families    : num [1:58] 191739 31939 16255 6800 2194080 ...
## $ hse_units   : num [1:58] 284367 43867 35492 12710 3445076 ...
## $ ave_fam_sz  : num [1:58] 3.61 3.59 2.94 2.98 3.58 3.63 2.94 2.77 3.02 3.74 ...
## $ vacant      : num [1:58] 29757 2634 8944 2652 203872 ...
## $ owner_occ   : num [1:58] 152828 22329 17472 6590 1544749 ...
## $ renter_occ  : num [1:58] 101782 18904 9076 3468 1696455 ...
## $ county_fips: chr [1:58] "06103" "06089" "06106" "06086" ...
## - attr(*, "spec")=
## .. cols(
## ..   ...1 = col_double(),
## ..   name = col_character(),
## ..   pop2012 = col_double(),
## ..   pop12_sqmi = col_double(),
## ..   white = col_double(),
## ..   black = col_double(),
## ..   ameri_es = col_double(),
## ..   asian = col_double(),
## ..   hawn_pi = col_double(),
## ..   hispanic = col_double(),
## ..   other = col_double(),
## ..   mult_race = col_double(),
## ..   males = col_double(),
## ..   females = col_double(),
## ..   med_age = col_double(),
```

```
## .. households = col_double(),
## .. families = col_double(),
## .. hse_units = col_double(),
## .. ave_fam_sz = col_double(),
## .. vacant = col_double(),
## .. owner_occ = col_double(),
## .. renter_occ = col_double(),
## .. county_fips = col_character()
## .. )
## - attr(*, "problems")=<externalptr>
```

```
typeof( demographics_data[["name"]])
```

```
## [1] "character"
```

```
typeof( demographics_data[["pop2012"]])
```

```
## [1] "double"
```

```
str(mortality_data)
```

```
## tibble [147,784 x 10] (S3: tbl_df/tbl/data.frame)
## $ Year          : num [1:147784] 2014 2014 2014 2014 2014 ...
## $ County        : chr [1:147784] "Alameda" "Alameda" "Alameda" "Alameda" ...
## $ Geography_Type : chr [1:147784] "Occurrence" "Occurrence" "Occurrence" "Occurrence" ...
## $ Strata         : chr [1:147784] "Total Population" "Age" "Age" "Age" ...
## $ Strata_Name    : chr [1:147784] "Total Population" "Under 1 year" "1-4 years" "5-14 years" ...
## $ Cause          : chr [1:147784] "ALL" "ALL" "ALL" "ALL" ...
## $ Cause_Desc     : chr [1:147784] "All causes (total)" "All causes (total)" "All causes (total)" "A
## $ Count          : num [1:147784] 9357 105 17 17 133 ...
## $ Annotation_Code: num [1:147784] 0 0 0 0 0 0 0 0 0 0 ...
## $ Annotation_Desc: chr [1:147784] "0" "0" "0" "0" ...
```

```
typeof(mortality_data[["County"]])
```

```
## [1] "character"
```

```
typeof(mortality_data[["Geography_Type"]])
```

```
## [1] "character"
```

```
typeof(mortality_data[["Cause"]])
```

```
## [1] "character"
```

```
typeof(mortality_data[["Count"]])
```

```
## [1] "double"
```

```
str(funding_data)
```

```
## tibble [53,592 x 7] (S3: tbl_df/tbl/data.frame)
## $ County                : chr [1:53592] "01 - Alameda" "01 - Alameda" "01 - Alameda" "01 - A
## $ Data Generation Date   : POSIXct[1:53592], format: "2013-10-14" "2013-10-14" ...
## $ OSHPD Project Status   : chr [1:53592] "In Review" "Pending Construction" "In Construction"
## $ Total Costs of OSHPD Projects: chr [1:53592] "$50,890,315.00" "$840,242,543.36" "$994,245,713.95"
## $ Number of OSHPD Projects : num [1:53592] 44 125 181 82 0 0 0 0 2 0 ...
## $ Collection of Counties  : chr [1:53592] "Bay Area Counties" "Bay Area Counties" "Bay Area Co
## $ Numeric_Cost           : num [1:53592] 5.09e+07 8.40e+08 9.94e+08 6.53e+07 0.00 ...
```

```
typeof(funding_data[["Total Costs of OSHPD Projects"]])
```

```
## [1] "character"
```

```
typeof(funding_data[["Numeric_Cost"]])
```

```
## [1] "double"
```

## demographics\_data

- The name column holds a variable of character string type, and seems to contain the name of counties. We may consider converting this into a Factor, will do so later on if we find such conversion to be useful.
- pop2012 is a numeric field containing the number of people of the named county, in 2012. We can perform computation such as mean calculations on this field, see below, so there isn't likely any need for conversion.

## mortality\_data

- County is a character string.
- Geography\_Type is a character string.
- Cause is a character string.
- Count is a number data type. It is the count of events.

## HCAI funding data

- County is a character string. However, there is also a number in it. eg "01 - Alameda". To join this data frame with the others, there is likely some manipulation needed to strip out the number part eg remove "01 -" and leave it with county names only
- Total Costs of OSHPD Projects was meant to be a numeric field, but it has dollar sign and commas, and so a string parsing to strip them out, and converted to numeric value, was done during the csv data import process above.

## Provide a basic description of the 5+ data elements

- Numeric: mean, median, range
- Character: unique values/categories
- Or any other descriptives that will be useful to the analysis

### demographics\_data

Code to count number of unique counties:

```
# Python style printf() function per
# https://stackoverflow.com/questions/13023274/how-to-do-printf-in-r
printf <- function(...) cat(sprintf(...))

# count number of unique name (ie counties)
uniq_counties = unique( demographics_data[["name"]]) %>% as.data.frame()
uniq_counties_count = count(uniq_counties)

printf( "The number of unique counties in the demographics data set was: %g",
        uniq_counties_count )
```

```
## The number of unique counties in the demographics data set was: 58
```

Code to find statistics of numerical data (population in 2012):

```
summary( demographics_data )
```

```
##           id           name           pop2012           pop12_sqmi
##  Min.      : 1.00   Length:58      Min.      : 1148   Min.      : 1.544
##  1st Qu.:15.25   Class :character  1st Qu.: 48492   1st Qu.: 25.887
##  Median :29.50   Mode  :character  Median : 180662   Median : 103.424
##  Mean   :29.50                      Mean   : 650129   Mean   : 665.061
##  3rd Qu.:43.75                      3rd Qu.: 645995   3rd Qu.: 333.485
##  Max.   :58.00                      Max.   :9904341   Max.   :17398.354
##           white           black           ameri_es           asian
##  Min.      : 881   Min.      : 0.0   Min.      : 44   Min.      : 7.0
##  1st Qu.: 38653   1st Qu.: 583.8   1st Qu.: 1102   1st Qu.: 672.5
##  Median : 137632   Median : 4083.0   Median : 2786   Median : 8782.0
##  Mean   : 369895   Mean   : 39639.2   Mean   : 6255   Mean   : 83810.5
##  3rd Qu.: 365881   3rd Qu.: 19117.8   3rd Qu.: 6397   3rd Qu.: 50296.0
##  Max.   :4936599   Max.   :856874.0   Max.   :72828   Max.   :1346865.0
##           hawn_pi           hispanic           other           mult_race
##  Min.      : 0.00   Min.      : 84   Min.      : 19   Min.      : 28
##  1st Qu.: 79.25   1st Qu.: 8964   1st Qu.: 3797   1st Qu.: 2111
##  Median : 350.50   Median : 44360   Median : 18380   Median : 7779
##  Mean   : 2489.41   Mean   : 241616   Mean   : 108920   Mean   : 31300
##  3rd Qu.: 1964.00   3rd Qu.: 226417   3rd Qu.: 109321   3rd Qu.: 35545
##  Max.   :26094.00   Max.   :4687889   Max.   :2140632   Max.   :438713
##           males           females           med_age           households
##  Min.      : 606   Min.      : 569   Min.      :29.60   Min.      : 497
```



```
## 1st Qu.: 24024 1st Qu.: 23597 1st Qu.:33.70 1st Qu.: 19041
## Median : 90108 Median : 90290 Median :37.05 Median : 70284
## Mean : 319273 Mean : 323037 Mean :38.49 Mean : 216853
## 3rd Qu.: 319545 3rd Qu.: 323048 3rd Qu.:43.08 3rd Qu.: 207712
## Max. :4839654 Max. :4978951 Max. :51.00 Max. :3241204
## families hse_units ave_fam_sz vacant
## Min. : 297 Min. : 1760 Min. :2.670 Min. : 827
## 1st Qu.: 13138 1st Qu.: 24679 1st Qu.:2.940 1st Qu.: 3362
## Median : 45541 Median : 76184 Median :3.245 Median : 8580
## Mean : 149008 Mean : 235863 Mean :3.212 Mean : 19010
## 3rd Qu.: 144280 3rd Qu.: 226459 3rd Qu.:3.493 3rd Qu.: 18544
## Max. :2194080 Max. :3445076 Max. :3.760 Max. :203872
## owner_occ renter_occ county_fips
## Min. : 357 Min. : 140 Length:58
## 1st Qu.: 13089 1st Qu.: 6080 Class :character
## Median : 39306 Median : 25140 Mode :character
## Mean : 121300 Mean : 95554
## 3rd Qu.: 120805 3rd Qu.: 84189
## Max. :1544749 Max. :1696455
```

```
# IQR for numeric data
```

```
Q1 = quantile( demographics_data[["pop2012"]], probs = 0.25, na.rm=T )
Median = median( demographics_data[["pop2012"]], na.rm=T )
Mean = mean( demographics_data[["pop2012"]], na.rm=T )
Q3 = quantile( demographics_data[["pop2012"]], probs = 0.75, na.rm=T )
Q1n = round( Q1[[1]], 2 )
Q3n = round( Q3[[1]], 2 )
```

```
printf( "The Mean for pop2012 in the demographics data set was found to be: %g",
        Mean )
```

```
## The Mean for pop2012 in the demographics data set was found to be: 650129
```

```
printf( "The interquartile range for pop2012 set was found to range from %g to %g",
        Q1n, Q3n )
```

```
## The interquartile range for pop2012 set was found to range from 48491.8 to 645995
```

- For the name column, it does not make much sense to talk about means or range, but we did find that our data set has 58 unique counties (ie, all counties of California is present in this data set)
- pop2012 has a mean of 650129, and an inter-quartile range of (48492, 645995)

## Mortality\_data Descriptions

```
#Summary of full dataset
```

```
summary(mortality_data)
```

```
##      Year      County      Geography_Type      Strata
```

```
## Min. :2014 Length:147784 Length:147784 Length:147784
## 1st Qu.:2015 Class :character Class :character Class :character
## Median :2017 Mode :character Mode :character Mode :character
## Mean :2017
## 3rd Qu.:2019
## Max. :2020
## Strata_Name Cause Cause_Desc Count
## Length:147784 Length:147784 Length:147784 Min. : 0.0
## Class :character Class :character Class :character 1st Qu.: 0.0
## Mode :character Mode :character Mode :character Median : 0.0
## Mean : 189.8
## 3rd Qu.: 41.0
## Max. :82816.0
## Annotation_Code Annotation_Desc
## Min. :0.000 Length:147784
## 1st Qu.:0.000 Class :character
## Median :0.000 Mode :character
## Mean :0.328
## 3rd Qu.:1.000
## Max. :2.000
```

```
# Unique characters: County, Geography Type, Causes
```

```
mortality_counties <- unique(mortality_data$County)
mortality_counties
```

```
## [1] "Alameda" "Alpine" "Amador" "Butte"
## [5] "Calaveras" "Colusa" "Contra Costa" "Del Norte"
## [9] "El Dorado" "Fresno" "Glenn" "Humboldt"
## [13] "Imperial" "Inyo" "Kern" "Kings"
## [17] "Lake" "Lassen" "Los Angeles" "Madera"
## [21] "Marin" "Mariposa" "Mendocino" "Merced"
## [25] "Modoc" "Mono" "Monterey" "Napa"
## [29] "Nevada" "Orange" "Placer" "Plumas"
## [33] "Riverside" "Sacramento" "San Benito" "San Bernardino"
## [37] "San Diego" "San Francisco" "San Joaquin" "San Luis Obispo"
## [41] "San Mateo" "Santa Barbara" "Santa Clara" "Santa Cruz"
## [45] "Shasta" "Sierra" "Siskiyou" "Solano"
## [49] "Sonoma" "Stanislaus" "Sutter" "Tehama"
## [53] "Trinity" "Tulare" "Tuolumne" "Ventura"
## [57] "Yolo" "Yuba"
```

```
mortality_geo_type <- unique(mortality_data$Geography_Type)
mortality_geo_type
```

```
## [1] "Occurrence" "Residence"
```

```
mortality_causes <- unique(mortality_data$Cause)
mortality_causes
```

```
## [1] "ALL" "ALZ" "CAN" "CLD" "DIA" "HOM" "HTD" "HYP" "INJ" "LIV" "NEP" "PAR"
## [13] "PNF" "STK" "SUI"
```

```
# Summary for Numeric Data - Count
summary(mortality_data$Count)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.0     0.0     0.0   189.8   41.0 82816.0
```

```
summary(mortality_data_raw$Count, na.rm=T)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
##      0.0     0.0    15.0   273.8   94.0 82816.0  45348
```

- There are 58 unique characters in the variable “County”.
- There are 2 unique characters in the variable “Geography\_Type”. If Geography\_Type is Resident, the death was of a CA resident but may have occurred out of state. If Geography\_Type is Occurrence, the death occurred in CA but the person may not have been a CA resident. Our research question asks us to focus on “occurrences”.
- There are 15 unique “Causes” of death. We are interested in chronic health conditions.
- The statistics for the numeric column Count on mortality events is a little tricky, as there are many na in this dataset. If we replaced na with 0, we find a mean of 189.8, a median is 0 and an IQR of 41. On the other hand, if we skip over entries with na, we find a mean of 273.8, a median of 15 and an IQR of 79. Finally, we observe that this column probably should have been “pivot wider” to segregate the various causes of disease vs Total, so the statistics obtained may not make a lot of sense. As we explore more on these datasets in future milestone we will develop a keener approach on how to handle these na's.

## HCAI Funding Data Descriptions

```
summary(funding_data)
```

```
##      County      Data Generation Date      OSHPD Project Status
## Length:53592    Min.   :2013-10-14 00:00:00.00    Length:53592
## Class :character 1st Qu.:2015-11-19 00:00:00.00    Class :character
## Mode  :character Median :2018-02-08 00:00:00.00    Mode  :character
##                  Mean   :2018-02-19 11:19:28.82
##                  3rd Qu.:2020-05-07 00:00:00.00
##                  Max.   :2022-08-11 00:00:00.00
## Total Costs of OSHPD Projects Number of OSHPD Projects Collection of Counties
## Length:53592      Min.   : 0.00      Length:53592
## Class :character  1st Qu.: 1.00      Class :character
## Mode  :character  Median : 6.00      Mode  :character
##                  Mean   : 27.94
##                  3rd Qu.: 23.00
##                  Max.   :1055.00
## Numeric_Cost
## Min.   :0.000e+00
## 1st Qu.:9.807e+04
## Median :2.824e+06
## Mean   :5.914e+07
## 3rd Qu.:2.845e+07
## Max.   :2.340e+09
```

- we find that the OSHPD Project costs has a mean value of \$59M and a median of \$2.8M