Group S: Milestone #2

PhuongTseng & Carri Beshears

2022-10-02

This is a team assignment; each team should complete and turn in a PDF created from an Rmd via Github. Please include code and output for the following components:

Description of dataset

What is the data source? (1-2 sentences on where the data is coming from, dates included, etc.) How does the data set relate to the group problem statement and question?

ANSWER: This project will use 3 data sets to identify 5 California counties for additional funding for healthcare facility improvement. First, information from the 2012 California county census will be reviewed to to isolate counties in rural areas, with residents of a high median age, and with a high proportion of renters over homeowners. Next, mortaility surveillance data will be used to determine mortality from chronic conditions will be aggregated by county for the years 2014-2020. Finally, HCAI data will be used to account for current healthcare funding by county from 2014-2020. All of these data sets are avaliable from the CA.gov open data portal.

Import statement

NOTE: Please use datasets available in the PHW251 Project Data github repoLinks 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.)

```r
#use dplyr::select specific columns
demog <-
  ca_county_demographic %>%
  dplyr::select(name, pop2012, pop12_sqmi, med_age, owner_occ, renter_occ)
#check first 2 records
head(demog, 2)
```

Document the import process

```
##     name pop2012 pop12_sqmi med_age owner_occ renter_occ
## 1:  Kern  851089   104.2829    30.7    152828     101782
## 2: Kings  155039   111.4274    31.1     22329      18904
```

```
#use dplyr::rename_with to make column names lower case
mortality <-
  ca_county_mortality %>%
  dplyr::rename_with( ~ tolower(gsub(" ", "_", .x, fixed = TRUE)))

head(mortality, 2)
```

```
##    year  county geography_type           strata      strata_name cause
## 1: 2014 Alameda     Occurrence Total Population Total Population   ALL
## 2: 2014 Alameda     Occurrence              Age     Under 1 year   ALL
##            cause_desc count annotation_code annotation_desc
## 1: All causes (total)  9357              NA
## 2: All causes (total)   105              NA
```

```
#same as above and substitute a dash "_" for space
construction <-
  hcai_healthcare_construction %>%
  dplyr::rename_with( ~ tolower(gsub(" ", "_", .x, fixed = TRUE)))
#see the last 2 records
tail(construction, 2)
```

```
##       county data_generation_date oshpd_project_status
## 1: 58 - Yuba           2022-08-11      In Construction
## 2: 58 - Yuba           2022-08-11           In Closure
##    total_costs_of_oshpd_projects number_of_oshpd_projects
## 1:                $3,756,441.72                         8
## 2:                             0                        0
##    collection_of_counties
## 1:
## 2:
```

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?

ANSWER: The 5 elements that will be used for this scenario start with 2 existing variables showing population per square mile (pop12_sqmi) and median age for the 2012 California county census (med_age).

```
typeof(ca_county_demographic$pop12_sqmi)
```

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

```
typeof(ca_county_demographic$med_age)
```

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

```
typeof(ca_county_demographic$owner_occ)
```

```
## [1] "integer"
```

```
typeof(ca_county_demographic$renter_occ)
```

```
## [1] "integer"
```

```
typeof(ca_county_mortality$Count)
```

```
## [1] "integer"
```

```
typeof(hcai_healthcare_construction$`Total Costs of OSHPD Projects`)
```

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

An additional variable will be created from 2 existing variables (owner_occ and renter_occ) to show the proportion of renters to homeowners. These are both currently numeric type variables and will not require a conversion. 2 additional variables will be added showing the county level mortality counts and project funding from raw data. These are both current stored at character type variables and will need to be converted to numeric type to aggregate and allow for analysis.

Provide a basic description of the 5+ data elements

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

```
summary(ca_county_demographic$pop12_sqmi)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##     1.544    25.887   103.424   665.061   333.485 17398.354
```

```
summary(ca_county_demographic$med_age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.60   33.70   37.05   38.49   43.08   51.00
```

```
summary(ca_county_demographic$owner_occ)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     357   13089   39306  121300  120804 1544749
```

```
summary(ca_county_demographic$renter_occ)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     140    6080   25140   95554   84189 1696455
```

Other required elements that will be used for analysis includes the aggregated data that will be created to determine county level information on mortality counts and project funding. These items are currently stored as character variables and will need to be converted prior to working with and moving them onto the main data set.