GroupS_Milestone3

PhuongTseng & Carri Beshears

2022-11-07

Scenario 1: Hospital Funding and Health Equity

You are a researcher in the California Department of Public Health Office of Health Equity (OHE). A policy has just been created to fund a public-private partnership for healthcare facility improvement in rural areas of California that have received minimal funding from the Department of Health Care Access and Information (HCAI) over the past 5 years. You are tasked with exploring and evaluating which 5 counties are the best targets for the development fund proposals. There are multiple components to this request.

First, OHE would like you to focus on rural areas, non-homeowners, and aging individuals as populations of interest in your analysis. Your task is to explore the California county census demographic dataLinks to an external site and begin to identify counties that share three common attributes:

1) low population per square mile `pop12_sqmi1`,
   According to the U.S. Census on population density https://www.census.gov/newsroom/blogs/random-samplings/2015/03/understanding-population-density.html, we will define low population per square mile as less 100 population per sq mile.

2) high median age `med_age`,
   According to https://www.ppic.org/publication/californias-population/ and the U.S. Census, California's median age in 2020 is 37.3, therefore, anything above 37 will be considered as above the median age.

3) a high proportion of renters vs. homeowners (you may need to create a new variable for this third criteria). We're defining high proportion of renters vs. homeowners as renters over the total population of renters and owners occupied household

Milestone 3 Criteria: Subset rows or columns as needed - 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 - If no new values are needed for final tables/graphs, please create 2 new variables anyway

```r
demog <-
  ca_county_demographic %>%
  dplyr::select(name, pop2012, pop12_sqmi, med_age, owner_occ, renter_occ) %>%
          mutate(pop12_sqmi1 = if_else(pop12_sqmi <= 100, "low", "not low"),
                 prop_rent_own = round((renter_occ/(renter_occ + owner_occ)),2),
                 high_p_renters =
                   renter_occ > owner_occ, high_med_age = med_age > 37) %>%
          rename(county = name)

#check first 2 records
head(demog, 2)
```

```
##    county pop2012 pop12_sqmi med_age owner_occ renter_occ pop12_sqmi1
## 1:   Kern  851089   104.2829    30.7    152828     101782     not low
## 2:  Kings  155039   111.4274    31.1     22329      18904     not low
##    prop_rent_own high_p_renters high_med_age
## 1:          0.40          FALSE        FALSE
## 2:          0.46          FALSE        FALSE
```

Clean variables needed for analysis (minimum 2)

- Examples
  - Recode invalid values
  - Handle missing fields
  - Recode categories
- If not needed for final analysis, please create at least 2 new variables anyway

```
#use dplyr::rename_with to make column names lower case
mortality <-
  ca_county_mortality %>%
  dplyr::rename_with(~ tolower(gsub(" ", "_", .x, fixed = TRUE))) %>%
  mutate(count = na_if(count, 0),
         annotation_code = na_if(annotation_code, 0),
         annotation_desc = na_if(annotation_desc, "NA"))

mortality$count[is.na(mortality$count)] <- 0
mortality$annotation_code[is.na(mortality$annotation_code)] <- 0
mortality$annotation_desc[is.na(mortality$annotation_desc)] <- "NA"


mortality2 <- mortality %>%
  group_by(county
           #,
           #geography_type,
           #strata,
           #strata_name,
           #cause,
           #cause_desc,
           #annotation_code,
           #annotation_desc
           ) %>%
  summarize(totalcount = sum(count)) %>%
  rename(countmortality = totalcount)


#mortality_wide <- mortality2 %>%
#  pivot_wider(names_from = year, values_from = count)

#same as above and substitute a dash "_" for space
hcai <-
  hcai_healthcare_construction %>%
  dplyr::rename_with( ~ tolower(gsub(" ", "_", .x, fixed = TRUE))) %>%
  mutate(collection_of_counties = na_if(collection_of_counties, "NA"))

library(lubridate)
```

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

## The following objects are masked from 'package:data.table':
##
```

```
##     hour, isoweek, mday, minute, month, quarter, second, wday, week,
##     yday, year

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

```
hcai$data_generation_date <- ymd(hcai$data_generation_date)
class(hcai$data_generation_date) #date
```

```
## [1] "Date"
```

```
hcai2 <- hcai %>% separate(county, c('county_code', 'county'), sep = " - ") %>%
  mutate(total_costs_of_oshpd_projects =
           as.character(gsub("[\\$,]", "", hcai$total_costs_of_oshpd_projects)))

hcai3 <-  hcai2 %>%
  group_by(county
           #,
           #geography_type,
           #strata,
           #strata_name,
           #cause,
           #cause_desc,
           #annotation_code,
           #annotation_desc
           ) %>%
  mutate(total_costs_of_oshpd_projects =
           as.numeric(total_costs_of_oshpd_projects)) %>%
  summarize(totalcosts = sum(total_costs_of_oshpd_projects))

merge_df <- merge(mortality2, demog, by="county")
merge_df2 <- merge(merge_df, hcai3, by = "county")
#hcai <- as.data.frame(hcai) %>%
#  separate(hcai$county, c("test", "test1"), " - ")

tail(merge_df2)
```

```
##      county countmortality pop2012 pop12_sqmi med_age owner_occ renter_occ
## 53  Trinity          10413   14063   4.384289    49.2      4284       1799
## 54   Tulare         311422  448724  92.738012    29.6     76586      53766
## 55 Tuolumne          62206   55331  24.304973    47.1     15471       6685
## 56  Ventura         597254  825977 444.788666    36.2    174168      92752
## 57     Yolo         124671  204322 199.657989    30.5     37416      33456
## 58     Yuba          70095   72822 113.153192    32.2     14468       9839
##    pop12_sqmi1 prop_rent_own high_p_renters high_med_age   totalcosts
## 53         low          0.30          FALSE         TRUE    243234641
## 54         low          0.41          FALSE        FALSE  19822918816
## 55         low          0.30          FALSE         TRUE    405954823
## 56     not low          0.35          FALSE        FALSE 107940103534
## 57     not low          0.47          FALSE        FALSE   7480514461
## 58     not low          0.40          FALSE        FALSE  22754058702
```

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

- Variable name
- Data type
- Description

```
library(kableExtra)
```

```
##
## Attaching package: ’kableExtra’

## The following object is masked from ’package:dplyr’:
##
##     group_rows
```

```
dictionary <- data.frame(
  columns = c(colnames(merge_df2)),
  type = c(
    "character" ,
    "numeric" ,
    "numeric",
    "numeric" ,
    "numeric",
    "numeric" ,
    "numeric",
    "character" ,
    "numeric" ,
    "boolean" ,
    "boolean",
    "numeric"
  ),
  description = c(
    "names of county",
    "count of mortality",
    "population in 2012",
    "population in 2012 per square mile",
    "median age",
    "owner occupied households",
    "renter occupied households",
    "population in 2012 per square mile with less than 100 persons per sqm",
    "proportion of renters versus owners",
    "high proportion of renters",
    "median age above 37 years old is considered as high median age",
    "total costs of oshpd project"))
```

```r
kable(dictionary)
```

columns	type	description
county	character	names of county
countmortality	numeric	count of mortality
pop2012	numeric	population in 2012
pop12_sqmi	numeric	population in 2012 per square mile
med_age	numeric	median age
owner_occ	numeric	owner occupied households
renter_occ	numeric	renter occupied households
pop12_sqmi1	character	population in 2012 per square mile with less than 100 persons per sqm
prop_rent_own	numeric	proportion of renters versus owners
high_p_renters	boolean	high proportion of renters
high_med_age	boolean	median age above 37 years old is considered as high median age
totalcosts	numeric	total costs of oshpd project

```r
#kable(head(merge_df2, 5), format = "html", caption = "Dataset")
```

```
kable(merge_df2)
```

county	countmortality	pop2012	pop12_sqmi	med_age	owner_occ	renter_occ	pop12_sqmi1	pro
Alameda	1037483	1534551	2062.402226	36.6	291242	253896	not low	
Alpine	167	1148	1.543841	46.4	357	140	low	
Amador	39388	38354	63.288340	48.2	10883	3686	low	
Butte	244224	222350	132.554757	37.1	50991	36627	not low	
Calaveras	41228	46212	44.582939	49.1	14520	4366	low	
Colusa	11178	21780	18.833988	33.5	4318	2738	low	
Contra Costa	807039	1067570	1405.326067	38.4	251904	123460	not low	
Del Norte	25924	28685	28.298164	39.0	6114	3793	low	
El Dorado	145457	182494	102.156840	43.5	51391	18832	not low	
Fresno	749101	944788	157.172588	30.7	158691	130700	not low	
Glenn	18880	28516	21.488749	35.3	6100	3700	low	
Humboldt	132542	136375	38.062105	37.3	30820	25211	low	
Imperial	104677	178091	39.744560	32.0	27465	21661	low	
Inyo	16996	18611	1.819773	45.5	5121	2928	low	
Kern	636963	851089	104.282870	30.7	152828	101782	not low	
Kings	81573	155039	111.427421	31.1	22329	18904	not low	
Lake	73044	65253	49.082334	45.0	17472	9076	low	
Lassen	17349	35039	7.422856	37.0	6590	3468	low	
Los Angeles	6978376	9904341	2423.264150	34.8	1544749	1696455	not low	
Madera	99921	153025	71.065672	33.1	27726	15591	low	
Marin	201166	255509	486.100489	44.5	64637	38573	not low	
Mariposa	14469	18455	12.613887	49.2	5227	2466	low	
Mendocino	80567	88094	25.083070	41.6	20601	14344	low	
Merced	162334	256841	129.897434	29.6	41196	34446	not low	
Modoc	7432	9791	2.329272	46.0	2786	1278	low	
Mono	2937	14418	4.604772	37.2	3228	2540	low	
Monterey	262748	420465	126.859300	33.0	64077	61869	not low	
Napa	123605	135855	172.308609	39.7	30597	18279	not low	
Nevada	99497	99951	102.564339	47.5	29890	11637	not low	
Orange	2170944	3054269	3822.423158	36.2	588313	404468	not low	
Placer	360882	356116	237.083491	40.3	94223	38404	not low	
Plumas	16422	20000	7.653217	49.5	6235	2742	low	
Riverside	1783174	2227789	305.044946	33.7	462212	224048	not low	
Sacramento	1247764	1432457	1441.219615	34.8	295482	218463	not low	
San Benito	27040	56501	40.634754	34.3	10927	5878	low	
San Bernardino	1550089	2062041	102.560224	31.7	383573	228045	not low	
San Diego	2313172	3137431	740.583699	34.7	591025	495840	not low	
San Francisco	616252	824334	17398.353736	38.5	123646	222165	not low	
San Joaquin	571080	688477	482.643869	32.7	127270	87737	not low	
San Luis Obispo	244105	271619	81.815416	39.4	60920	41096	low	
San Mateo	487818	726677	1591.217045	39.2	153110	104727	not low	
Santa Barbara	331556	423800	154.042992	33.7	74827	67277	not low	
Santa Clara	1092595	1819137	1401.071327	36.2	348298	255906	not low	
Santa Cruz	175223	262470	587.522944	36.8	54229	40126	not low	
Shasta	240864	178831	46.480517	41.8	45277	25069	low	
Sierra	420	3226	3.353291	51.0	1065	417	low	
Siskiyou	51403	45200	7.120891	46.8	12629	6876	low	
Solano	343547	418187	470.005058	36.9	89648	52110	not low	
Sonoma	422984	487061	306.323820	39.8	112280	73545	not low	
Stanislaus	485852	518549	342.538842	32.9	99364	65816	not low	
Sutter	66912	95619	157.125955	34.6	19212	12225	not low	
Tehama	58742	63757	21.523312	39.5	15363	8404	low	
Trinity	10413	14063	4.384289	49.2	4284	1799	low	
Tulare	311422	448724	92.738012	29.6	76586	53766	low	
Tuolumne	62206	55331	24.304973	47.1	15471	6685	low	
Ventura	597254	825977	444.788666	36.2	174168	92752	not low	
Yolo	124671	204322	199.657989	30.5	37416	33456	not low	
Yuba	79895	72833	113.153102	33.3	14468	9839	not low	

```
library(kableExtra)
kable(summary(merge_df2))
```

	county	countmortality	pop2012	pop12_sqmi	med_age	owner_occ	renter
	Length:58	Min. : 167	Min. : 1148	Min. : 1.544	Min. :29.60	Min. : 357	Min.
	Class :character	1st Qu.: 43772	1st Qu.: 48492	1st Qu.: 25.887	1st Qu.:33.70	1st Qu.: 13089	1st Q
	Mode :character	Median : 139000	Median : 180662	Median : 103.424	Median :37.05	Median : 39306	Media
	NA	Mean : 483641	Mean : 650129	Mean : 665.061	Mean :38.49	Mean : 121300	Mean
	NA	3rd Qu.: 487326	3rd Qu.: 645995	3rd Qu.: 333.485	3rd Qu.:43.08	3rd Qu.: 120804	3rd Q
	NA	Max. :6978376	Max. :9904341	Max. :17398.354	Max. :51.00	Max. :1544749	Max.

```
summary(merge_df2)
```

```
##     county          countmortality      pop2012          pop12_sqmi
##  Length:58          Min.   :    167   Min.   :   1148   Min.   :    1.544
##  Class :character   1st Qu.:  43772   1st Qu.:  48492   1st Qu.:   25.887
##  Mode  :character   Median : 139000   Median : 180662   Median :  103.424
##                     Mean   : 483641   Mean   : 650129   Mean   :  665.061
##                     3rd Qu.: 487326   3rd Qu.: 645995   3rd Qu.:  333.485
##                     Max.   :6978376   Max.   :9904341   Max.   :17398.354
##     med_age        owner_occ         renter_occ      pop12_sqmi1
##  Min.   :29.60   Min.   :    357   Min.   :    140   Length:58
##  1st Qu.:33.70   1st Qu.:  13089   1st Qu.:   6080   Class :character
##  Median :37.05   Median :  39306   Median :  25140   Mode  :character
##  Mean   :38.49   Mean   : 121300   Mean   :  95554
##  3rd Qu.:43.08   3rd Qu.: 120804   3rd Qu.:  84189
##  Max.   :51.00   Max.   :1544749   Max.   :1696455
##  prop_rent_own    high_p_renters  high_med_age      totalcosts
##  Min.   :0.2300   Mode :logical   Mode :logical   Min.   :0.000e+00
##  1st Qu.:0.3400   FALSE:56        FALSE:29        1st Qu.:1.581e+09
##  Median :0.3850   TRUE :2         TRUE :29        Median :9.291e+09
##  Mean   :0.3833                                   Mean   :5.465e+10
##  3rd Qu.:0.4275                                   3rd Qu.:3.961e+10
##  Max.   :0.6400                                   Max.   :5.627e+11
```

PDF that is professionally prepared for presentation

- Each part of the milestone is clearly on one page (use to push to a new page)

- Only the necessary information is outputted (you should suppress, for example, entire data frame outputs)
- Use of headers and sub headers to create an organized document