R Project Milestone 3

Moyra Rasheed, Courtney Coon, Jarett Maycott

2022-10-28

Demographic data

```
#subset demographic dataset to include only needed columns
demo_sub<-select(demo_data, c("pop12_sqmi", "name", "med_age","renter_occ", "owner_occ"))
head(demo_sub)
```

```
##    pop12_sqmi        name med_age renter_occ owner_occ
## 1  104.282870        Kern    30.7     101782    152828
## 2  111.427421       Kings    31.1      18904     22329
## 3   49.082334        Lake    45.0       9076     17472
## 4    7.422856      Lassen    37.0       3468      6590
## 5 2423.264150 Los Angeles    34.8    1696455   1544749
## 6   71.065672      Madera    33.1      15591     27726
```

```
#categorize median age as low, medium, high
sum_med_age<-summary(demo_sub$med_age)
sum_med_age
```

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

```
demo_sub<-demo_sub%>%
  mutate(med_age_CAT= case_when(
    med_age < as.numeric(sum_med_age[2]) ~ "Low",
    med_age < as.numeric(sum_med_age[5]) ~ "Medium",
    TRUE ~ "High"))%>%
  mutate(med_age_CAT = factor(med_age_CAT, levels = c("Low", "Medium", "High")))

#categorize pop12_sqmi as low medium high
sum_pop12_sqmi<-summary(demo_sub$pop12_sqmi)
sum_pop12_sqmi
```

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

```r
demo_sub<-demo_sub%>%
  mutate(pop12_sqmi_CAT= case_when(
    pop12_sqmi < as.numeric(pop12_sqmi[2]) ~ "Low",
    pop12_sqmi < as.numeric(pop12_sqmi[5]) ~ "Medium",
    TRUE ~ "High"))%>%
  mutate(pop12_sqmi_CAT = factor(pop12_sqmi_CAT,
    levels = c("Low", "Medium", "High")))

#create renter:owner variable
demo_sub<-demo_sub%>%
  mutate(renter_owner_ratio=renter_occ/owner_occ)

#categorize renter_owner_ratio as low medium high
sum_renter_owner_ratio<-summary(demo_sub$renter_owner_ratio)
sum_renter_owner_ratio
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3007  0.5212  0.6272  0.6472  0.7382  1.7968
```

```r
demo_sub<-demo_sub%>%
  mutate(renter_owner_ratio_CAT= case_when(
    renter_owner_ratio < as.numeric(renter_owner_ratio[2]) ~ "Low",
    renter_owner_ratio < as.numeric(renter_owner_ratio[5]) ~ "Medium",
    TRUE ~ "High"))%>%
  mutate(renter_owner_ratio_CAT = factor(renter_owner_ratio_CAT,
    levels = c("Low", "Medium", "High")))

demo_sub<-select(demo_sub, c("pop12_sqmi_CAT", "name", "med_age_CAT","renter_owner_ratio_CAT"))
```

Mortality data

```r
#filter by 'strata_name' as "total_population'
mort_sub<-mort_data%>%
  filter(strata_name=="Total Population")

#replace NAs with zeros
mort_sub <- mort_sub %>% mutate(count = ifelse(is.na(count), 0, count))

#filter non-chronic diseases
unique(mort_data$cause_desc)
```

```
##  [1] "All causes (total)"
##  [2] "Alzheimer’s disease"
##  [3] "Malignant neoplasms"
##  [4] "Chronic lower respiratory diseases"
##  [5] "Diabetes mellitus"
##  [6] "Assault (homicide)"
##  [7] "Diseases of heart"
##  [8] "Essential hypertension and hypertensive renal disease"
##  [9] "Accidents (unintentional injuries)"
## [10] "Chronic liver disease and cirrhosis"
## [11] "Nephritis, nephrotic syndrome and nephrosis"
## [12] "Parkinson’s disease"
## [13] "Influenza and pneumonia"
## [14] "Cerebrovascular diseases"
## [15] "Intentional self-harm (suicide)"
```

```r
  ##remove "all cause", "assault", "accidents", "influenza" and "self-harm"
mort_sub<-mort_sub%>%
  filter(cause_desc %in% c("Alzheimer's disease", "Malignant neoplasms",
                           "Chronic lower respiratory diseases","Diabetes mellitus",
                           "Diseases of heart", "Essential hypertension and hypertensive renal disease"
                           "Chronic liver disease and cirrhosis", "Nephritis, nephrotic syndrome and nep
                           "Parkinson's disease", "Cerebrovascular diseases"))
unique(mort_sub$cause_desc)
```

```
##  [1] "Alzheimer’s disease"
##  [2] "Malignant neoplasms"
##  [3] "Chronic lower respiratory diseases"
##  [4] "Diabetes mellitus"
##  [5] "Diseases of heart"
##  [6] "Essential hypertension and hypertensive renal disease"
##  [7] "Chronic liver disease and cirrhosis"
##  [8] "Nephritis, nephrotic syndrome and nephrosis"
##  [9] "Parkinson’s disease"
## [10] "Cerebrovascular diseases"
```

```r
#summarize chronic death mortality by county
mort_sub_grouped<-mort_sub%>%
  group_by(county)%>%
  summarize(summed_chronic_dis_mort=sum(count))

#make summed_chronic_dis_mort categorical
sum_summed_chronic_dis_mort<-summary(mort_sub_grouped$summed_chronic_dis_mort)
sum_summed_chronic_dis_mort
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    4274   13413   46904   47785  698885
```

```
mort_sub_grouped<-mort_sub_grouped%>%
  mutate(sum_summed_chronic_dis_mort_CAT= case_when(
    summed_chronic_dis_mort < as.numeric(sum_summed_chronic_dis_mort[2]) ~ "Low",
    summed_chronic_dis_mort < as.numeric(sum_summed_chronic_dis_mort[5]) ~ "Medium",
    TRUE ~ "High"))%>%
  mutate(sum_summed_chronic_dis_mort_CAT = factor(sum_summed_chronic_dis_mort_CAT,
    levels = c("Low", "Medium", "High")))
```

Healthcare data

```
#filter data to only include projects with 'project status' as 'in closure'
healthcare_sub<-healthcare_data%>%
  filter(oshpd_project_status=="In Closure")

#filter out 2013 data
healthcare_sub<-healthcare_sub%>%
  filter(data_generation_date > "2014-01-01")

#summarize 'total_cost' of projects 'in closure' over the 5 years (2014-2020)
healthcare_sub_grouped<-healthcare_sub%>%
  group_by(county)%>%
  summarize(summed_number_oshpd_projects=sum(number_of_oshpd_projects))

#make summed_number_oshpd_projects categorical
sum_summed_number_oshpd_projects<-summary(healthcare_sub_grouped$summed_number_oshpd_projects)
sum_summed_number_oshpd_projects
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   252.2   880.0  3737.7  2951.0 67036.0
```

```
healthcare_sub_grouped<-healthcare_sub_grouped%>%
  mutate(summed_number_oshpd_projects_CAT= case_when(
    summed_number_oshpd_projects < as.numeric(sum_summed_number_oshpd_projects[2]) ~ "Low",
    summed_number_oshpd_projects < as.numeric(sum_summed_number_oshpd_projects[5]) ~ "Medium",
    TRUE ~ "High"))%>%
  mutate(summed_number_oshpd_projects_CAT = factor(summed_number_oshpd_projects_CAT,
    levels = c("Low", "Medium", "High")))
```

Merge the 3 datasets

```
#merge data sets by county
merged_data<-full_join(mort_sub_grouped, healthcare_sub_grouped, by="county")
merged_data<-full_join(merged_data, demo_sub, by=c("county"= "name"))
merged_data <- subset( merged_data, select = -summed_chronic_dis_mort )
merged_data <- subset( merged_data, select = -summed_number_oshpd_projects )
head(merged_data)
```

```
## # A tibble: 6 x 6
##   county    sum_summed_chronic_dis_~ summed_number_o~ pop12_sqmi_CAT med_age_CAT
##   <chr>     <fct>                    <fct>            <fct>          <fct>
## 1 Alameda   High                     High             Medium         Medium
## 2 Alpine    Low                      Low              Low            High
## 3 Amador    Low                      Medium           Low            High
## 4 Butte     Medium                   Medium           Medium         Medium
## 5 Calaveras Low                      Low              Low            High
## 6 Colusa    Low                      Low              Low            Low
## # ... with 1 more variable: renter_owner_ratio_CAT <fct>
```

Data Dictionary

variable 1: “county”

The county the data comes from

variable 2: “summed_chronic_dis_mort_CAT”

This is the summed number of mortality cases from chronic diseases in each county from 2014-2020. It was categorized into low, medium and high based on whether density was below the 1st quantile (low), between the 1st and 3rd quantile (medium), or above the 3rd quantile (high) for the state of California.

variable 3: “summed_number_oshped_projects_CAT”

This is the total number of closed projects per county from 1/1/2014 through 8/11/2022. It was categorized into low, medium and high based on whether density was below the 1st quantile (low), between the 1st and 3rd quantile (medium), or above the 3rd quantile (high) for the state of California.

variable 4: “pop12_sqmi_CAT”

This is population density at the county level categorized into low, medium and high based on whether density was below the 1st quantile (low), between the 1st and 3rd quantile (medium), or above the 3rd quantile (high) for the state of California.

variable 5: “med_age_CAT”

This is median age of residents at the county level categorized into low, medium and high based on whether median age was below the 1st quantile (low), between the 1st and 3rd quantile (medium), or above the 3rd quantile (high) for the state of California.

variable 6: “renter_owner_ratio_CAT”

This variable was created by dividing the number of renters by the number of owners in a county and then categorizing them into low, medium and high. The ratio was categorized as low if it was below the 1st quantile, medium if it was between 1st and 3rd quantile, or high if it was above the 3rd quantil as compared to other counties in the state of California.

Table with descriptive stats for variables in data dictionary

```
# df<-data.frame(
#   categories=c("Low", "Medium", "High"),
#   Chronic_disease_mortality=as.numeric(summary(merged_data$sum_summed_chronic_dis_mort_CAT)),
#   Number_oshpd_projects=as.numeric(summary(merged_data$summed_number_oshpd_projects_CAT)),
#   Number_oshpd_projects=as.numeric(summary(merged_data$summed_number_oshpd_projects_CAT)),
#   )
# df

library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'

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

```
table<-merged_data%>%
 rowwise() %>%
 mutate(number_lows= sum(c_across(2:6) == "Low", na.rm = TRUE),
        number_mediums= sum(c_across(2:6) == "Medium", na.rm = TRUE),
        temp_rank=(number_lows*2)+number_mediums
        )%>%
  ungroup()%>%
  arrange(desc(temp_rank))%>%
  select(-c(number_lows, number_mediums,temp_rank))
head(table)
```

```
## # A tibble: 6 x 6
##   county     sum_summed_chronic_dis~ summed_number_o~ pop12_sqmi_CAT med_age_CAT
##   <chr>      <fct>                   <fct>            <fct>          <fct>
## 1 Colusa     Low                     Low              Low            Low
## 2 Glenn      Low                     Low              Low            Medium
## 3 Imperial   Medium                  Low              Low            Low
## 4 Mono       Low                     Low              Low            Medium
## 5 San Benito Low                     Low              Low            Medium
## 6 Alpine     Low                     Low              Low            High
## # ... with 1 more variable: renter_owner_ratio_CAT <fct>
```

```
#     kable(
#       digits=1,
#       col.names = c("County","Chronic Disease Mortality","Number of closed OSHPD projects",
#                    "Population Density", "Median Age", "Renter:Owner ratio"),
#       caption="Relative levels of",
#       booktabs=TRUE,
#       align='lccccc',
#       escape=FALSE)
# table
```