R Project Milestone 3

Moyra Rasheed, Courtney Coon, Jarett Maycott

2022-10-28

Demographic data

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

#categorize median age: younger is lower priority, older is high priority
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 
```

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

#categorize pop12_sqmi with low density (rural) as high priority
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]) ~ "High priority",
    pop12_sqmi < as.numeric(pop12_sqmi[5]) ~ "Medium priority",
    TRUE ~ "Low priority"))%>%
  mutate(pop12_sqmi_CAT = factor(pop12_sqmi_CAT,
    levels = c("Low priority", "Medium priority", "High priority")))

#create % renter (of all households) variable
demo_sub<-demo_sub%>%
  mutate(renter_ratio=renter_occ/(owner_occ+renter_occ))
```

```
#categorize renter_ratio with higher ratio being a higher priority
sum_renter_ratio<-summary(demo_sub$renter_ratio)
sum_renter_ratio
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2312  0.3426  0.3854  0.3833  0.4247  0.6424
```

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

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

Mortality data

```
#filter by 'strata_name' as "total_population'
mort_sub<-mort_data%>%
  filter(strata_name=="Total Population")
#Note: We kept occurence and residence data because it's a metric of overall
#population load at the hospital

#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)"
```

```
#Note: We removed all non-chronic diseases: "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 nephrosi
                     "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"
```

```
#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 with higher counts of chronic disease
#in hospitals being higher priority
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(summed_chronic_dis_mort_CAT= case_when(
    summed_chronic_dis_mort < as.numeric(sum_summed_chronic_dis_mort[2]) ~ "Low priority",
    summed_chronic_dis_mort < as.numeric(sum_summed_chronic_dis_mort[5]) ~ "Medium priority",
    TRUE ~ "High priority"))%>%
  mutate(summed_chronic_dis_mort_CAT = factor(summed_chronic_dis_mort_CAT,
    levels = c("Low priority", "Medium priority", "High priority")))
```

Healthcare data

```
#filter data to only include projects with 'project status' as "Pending
#Construction," "In Construction" or "In Closure"
healthcare_sub<-healthcare_data%>%
  filter(oshpd_project_status!="In Review")
#Note: We only filtered out "in review" because all other projects with other
#statuses were guaranteed to be completed and we wanted to prioritize locations
#with the fewest projects that were at any stage of completion

#filter out 2013 data
healthcare_sub<-healthcare_sub%>%
  filter(data_generation_date > "2015-01-01")

#summarize 'total_cost' of projects 'in closure' over the 5 years (2015-2020)
#FIX: PRIORITIZE LOW COST AND LOW NUMBER OF PROJECTS
healthcare_sub_grouped<-healthcare_sub%>%
  group_by(county)%>%
  summarize(summed_total_cost=sum(total_costs_of_oshpd_projects))

#make the total cost of all oshpd projects categorical where counties with higher
#total costs are lower priority
s_summed_total_cost<-summary(healthcare_sub_grouped$summed_total_cost)
s_summed_total_cost
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.000e+00 1.112e+09 5.962e+09 3.389e+10 1.968e+10 3.285e+11
```

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

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 = -c(summed_chronic_dis_mort, summed_total_cost))
head(merged_data)
```

```
## # A tibble: 6 x 6
##   county    summed_chronic_dis_mort~ summed_total_co~ pop12_sqmi_CAT med_age_CAT
##   <chr>     <fct>                    <fct>            <fct>          <fct>
## 1 Alameda   High priority            Low priority     Medium priori~ Medium pri~
## 2 Alpine    Low priority             High priority    High priority  High prior~
## 3 Amador    Low priority             High priority    High priority  High prior~
## 4 Butte     Medium priority          Medium priority  Medium priori~ Medium pri~
## 5 Calaveras Low priority             High priority    High priority  High prior~
## 6 Colusa    Low priority             High priority    High priority  Low priori~
## # ... with 1 more variable: renter_ratio_CAT <chr>
```

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 2015-2020. It was categorized into low, medium and high priority based on whether density was below the 1st quantile (low priority), between the 1st and 3rd quantile (medium priority), or above the 3rd quantile (high priority) because we wanted to priority counties with higher burdens of chronic diseases.

variable 3: "summed_total_cost_CAT"

This is the total cost of all projects "Pending Construction," "In Construction," or "In Closure" from 1/1/2015 through 8/11/2022 (most recent data). It was categorized into low, medium and high priority based on whether summed total costs were below the 1st quantile (high priority), between the 1st and 3rd quantile (medium priority), or above the 3rd quantile (low priority) because we wanted to prioritize counties that had received less funding for oshpd projects.

variable 4: "pop12_sqmi_CAT"

This is population density at the county level categorized into low, medium and high priority based on whether density was below the 1st quantile (high priority), between the 1st and 3rd quantile (medium priority), or above the 3rd quantile (low priority) because we wanted to prioritize rural counties i.e., those with lower population densities.

variable 5: "med_age_CAT"

This is median age of residents at the county level categorized into low, medium and high priority based on whether median age was below the 1st quantile (low priority), between the 1st and 3rd quantile (medium priority), or above the 3rd quantile (high priority) because we wanted to prioritize counties with more elderly populations.

variable 6: "renter_ratio_CAT"

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

Table with descriptive stats for variables in data dictionary

```
library(kableExtra)
```

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

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

```
table<-merged_data%>%
 rowwise() %>%
 mutate(number_highs= sum(c_across(2:6) == "High priority", na.rm = TRUE),
        number_mediums= sum(c_across(2:6) == "Medium priority", na.rm = TRUE),
        temp_rank=(number_highs*2)+number_mediums
        )%>%
  ungroup()%>%
  arrange(desc(temp_rank))%>%
  select(-c(number_highs, number_mediums,temp_rank))%>%
  slice(1:10)
head(table)
```

```
## # A tibble: 6 x 6
##   county    summed_chronic_dis_mort~ summed_total_co~ pop12_sqmi_CAT med_age_CAT
##   <chr>     <fct>                    <fct>            <fct>          <fct>
## 1 Tuolumne  Medium priority          High priority    High priority  High prior~
## 2 Alpine    Low priority             High priority    High priority  High prior~
## 3 Amador    Low priority             High priority    High priority  High prior~
## 4 Calaveras Low priority             High priority    High priority  High prior~
## 5 El Dorado Medium priority          Medium priority  High priority  High prior~
## 6 Inyo      Low priority             High priority    High priority  High prior~
## # ... with 1 more variable: renter_ratio_CAT <chr>
```

```
kable(table,
       col.names = c("County","Chronic disease mortality burden",
                     "Previous spending on projects",
                     "Population density", "Median age of population",
                     "% population that are renters"),
       caption="Top 10 Counties ranked by need for oshpd projects.",
       booktabs=TRUE,
       align='lccccc')%>%
  kable_styling(latex_options="scale_down")
```

```
 table
```

```
## # A tibble: 10 x 6
##    county    summed_chronic_dis_mor~ summed_total_co~ pop12_sqmi_CAT med_age_CAT
##    <chr>     <fct>                   <fct>            <fct>          <fct>
##  1 Tuolumne  Medium priority         High priority    High priority  High prior~
```

Table 1: Top 10 Counties ranked by need for oshpd projects.

County	Chronic disease mortality burden	Previous spending on projects	Population density	Median age of population	% population that are renters
Tuolumne	Medium priority	High priority	High priority	High priority	Low priority
Alpine	Low priority	High priority	High priority	High priority	Low priority
Amador	Low priority	High priority	High priority	High priority	Low priority
Calaveras	Low priority	High priority	High priority	High priority	Low priority
El Dorado	Medium priority	Medium priority	High priority	High priority	Low priority
Inyo	Low priority	High priority	High priority	High priority	Low priority
Lake	Medium priority	Medium priority	High priority	High priority	Low priority
Mariposa	Low priority	High priority	High priority	High priority	Low priority
Nevada	Medium priority	Medium priority	High priority	High priority	Low priority
Plumas	Low priority	High priority	High priority	High priority	Low priority

```
##  2 Alpine    Low priority           High priority    High priority  High prior~
##  3 Amador    Low priority           High priority    High priority  High prior~
##  4 Calaveras Low priority           High priority    High priority  High prior~
##  5 El Dorado Medium priority        Medium priority  High priority  High prior~
##  6 Inyo      Low priority           High priority    High priority  High prior~
##  7 Lake      Medium priority        Medium priority  High priority  High prior~
##  8 Mariposa  Low priority           High priority    High priority  High prior~
##  9 Nevada    Medium priority        Medium priority  High priority  High prior~
## 10 Plumas    Low priority           High priority    High priority  High prior~
## # ... with 1 more variable: renter_ratio_CAT <chr>
```