R Project Milestone 3

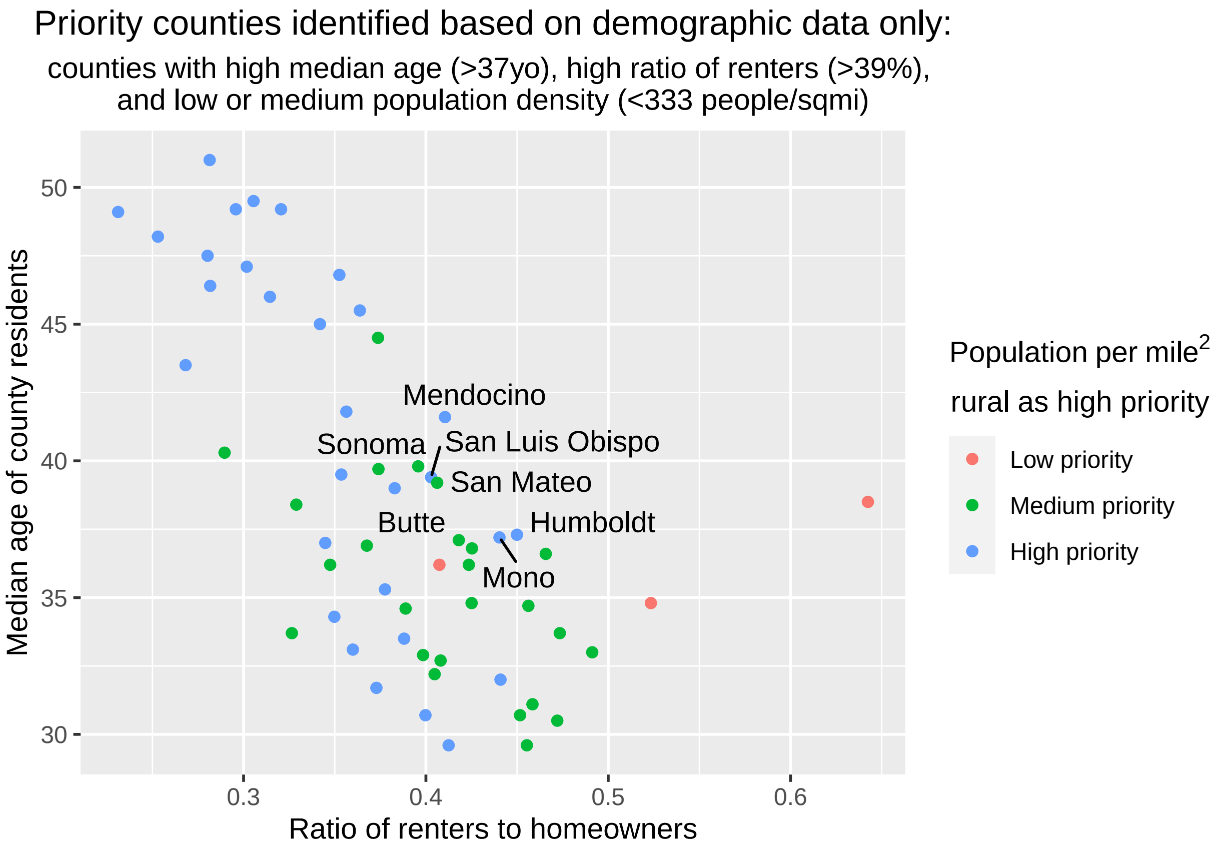
Moyra Rasheed, Courtney Coon, Jarett Maycott

2022-11-26

Visualization 1:

Plot 1: As a first step, we plotted demographic information only (population per square mile, renter to homeowner ratio, and median age of the residents of the county) and highlighted the counties that would be the highest priority if we were only looking at the demographic data. We prioritized low density, higher percentage of renters, and older populations.

```r
## renter ratio median = 39%
## median age median = 37.05
## population density 1st quantile (low cutoff) = 25.887
## population density 3rd quantile (high cutoff) = 333.485
ggplot(data = merged_data, aes(x = renter_ratio, y = med_age)) +
geom_point(data = merged_data, aes(x = renter_ratio, y = med_age,
                                   color = pop12_sqmi_CAT)) +
geom_text_repel(aes(label=ifelse((med_age > 37 & renter_ratio > 0.39
    & (pop12_sqmi_CAT=="High priority"| pop12_sqmi_CAT=="Medium priority")),
    county, "")))+
  labs(title = "Priority counties identified based on demographic data only:",
subtitle = "counties with high median age (>37yo), high ratio of renters (>39%),
and low or medium population density (<333 people/sqmi)",
       x = "Ratio of renters to homeowners",
       y = "Median age of county residents",
       color =
         bquote(atop(Population~per~mile^{"2"}, "rural as high priority")))+
  theme(plot.title=element_text(hjust=0.5),
        plot.subtitle=element_text(hjust=0.5))
```


Priority counties identified based on demographic data only:
counties with high median age (>37yo), high ratio of renters (>39%),
and low or medium population density (<333 people/sqmi)
Median age of county residents
50
45
40
35
30
0.3
0.4
0.5
0.6
Ratio of renters to homeowners
Mendocino
Sonoma
San Luis Obispo
San Mateo
Butte
Humboldt
Mono
Population per mile^2
rural as high priority
Low priority
Medium priority
High priority

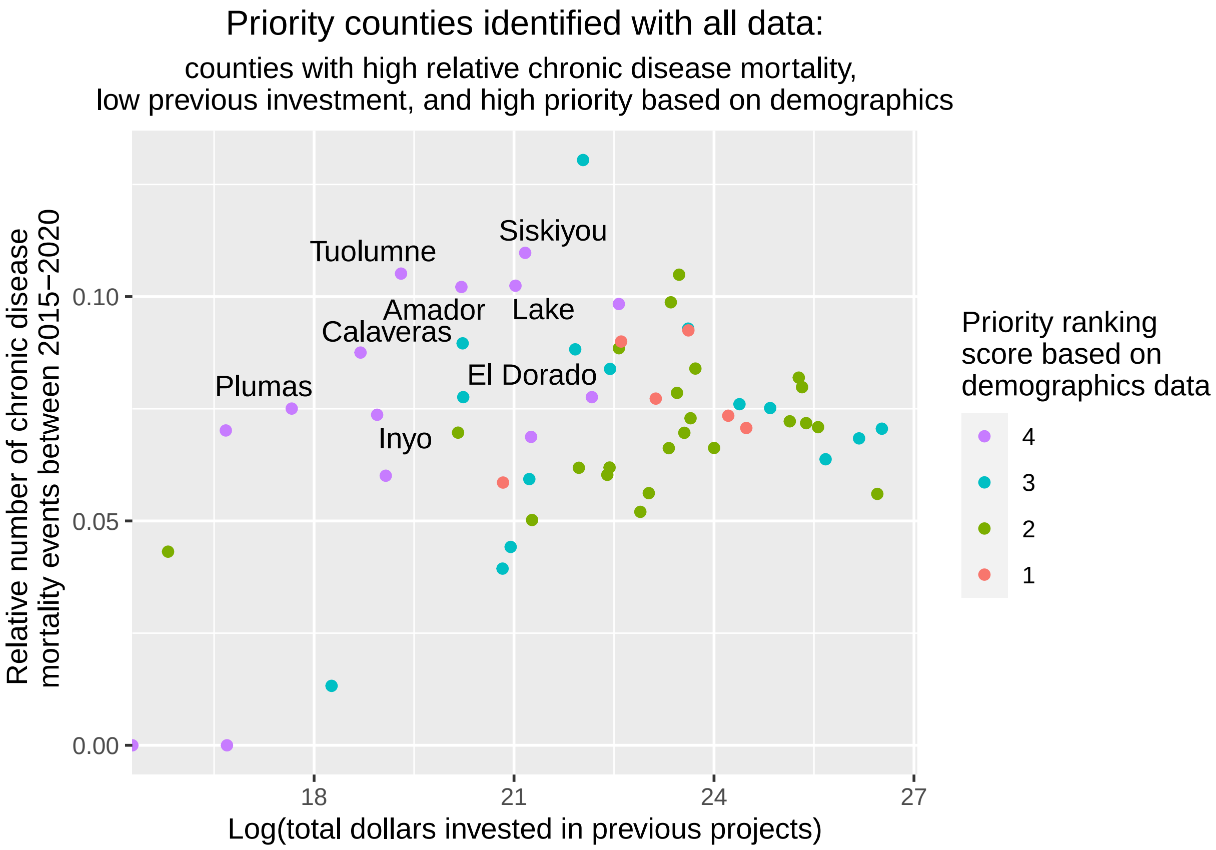
Visualization 2:

Plot 2: Here we integrated demographic data from Plot 1 with additional data: the total dollars invested in prior projects (with higher priority for less money previously invested) and number of chronic disease mortality events between 2015-2020, relativized by total population (with higher priority being higher relative levels of chronic disease mortality). We used demographic data from Plot 1 to create a new ranked variable which we used to color code data in Plot 2. Counties highlighted in Plot 1 were ranked more highly in Plot 2. Thus this Plot includes all 5 variables of interest to identify the counties that require greater funding. Note that money invested in previous projects has been logged to make the plot easier to read.

```
# make data set with continuous data and ranking factor for the demographic
# data in the first figure
second_fig_data_temp<-merged_data%>%
  select(c("county", "pop12_sqmi_CAT", "med_age_CAT", "renter_ratio_CAT"))%>%
 rowwise() %>%
 mutate(number_highs= sum(c_across(2:4) == "High priority", na.rm = TRUE),
        number_mediums= sum(c_across(2:4) == "Medium priority", na.rm = TRUE),
        demo_rank=(number_highs*2)+number_mediums
        )%>%
  ungroup()%>%
  select(c("county", "demo_rank"))

second_fig_data_final<-full_join(second_fig_data_temp, merged_data, by="county")
## summary(second_fig_data_final$relative_chronic_dis_mort)

# make the figure
## relative chronic disease mortality median = 0.07213
## log(relative chronic disease mortality median) = log(0.07213) = -2.629285
## summed total cost median = 5961782208
## log(summed total cost median) = log(5961782208) = 22.50864
ggplot(data = second_fig_data_final,
       aes(y = relative_chronic_dis_mort, x = log(summed_total_cost))) +
geom_point(data = second_fig_data_final,
           aes(y = relative_chronic_dis_mort, x = log(summed_total_cost),
                                   color = as.factor(demo_rank))) +
guides(color = guide_legend(reverse=TRUE))+
 geom_text_repel(aes(label=ifelse(
   (relative_chronic_dis_mort >= 0.07213 & summed_total_cost<=5961782208
    & demo_rank >3), county, "")), max.overlaps = Inf)+
labs(title = "Priority counties identified with all data:",
 subtitle = "counties with high relative chronic disease mortality,
low previous investment, and high priority based on demographics",
        x = "Log(total dollars invested in previous projects)",
        y =
"Relative number of chronic disease \n mortality events between 2015-2020",
        color = "Priority ranking \nscore based on \ndemographics data") +
   theme(plot.title=element_text(hjust=0.5),
         plot.subtitle=element_text(hjust=0.5))
```


Priority counties identified with all data:
counties with high relative chronic disease mortality,
low previous investment, and high priority based on demographics
Relative number of chronic disease mortality events between 2015−2020
0.10
0.05
0.00
Log(total dollars invested in previous projects)
18
21
24
27
Siskiyou
Tuolumne
Amador
Lake
Calaveras
El Dorado
Plumas
Inyo
Priority ranking score based on demographics data
4
3
2
1

Visualization 3:

Table 1: We used a table as a different way to organize the 5 variables of interest. We categorized each variable and then ranked each as high, medium, or low priority. We then created a new variable where “high priority” variables were given two points, “medium priority” variables given 1 point, and “low priority” variables given 0 points for each county. Then Counties are ranked by number of points. Most of the Counties highlighted in Plot/Visualization 2 were also the most highly ranked in the table.

```
library(kableExtra)
```

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

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

```
table_col_order <- c("county", "summed_total_cost", "pop12_sqmi",
                     "med_age", "renter_ratio",
                     "relative_chronic_dis_mort", "med_age_CAT",
                     "summed_total_cost_CAT", "pop12_sqmi_CAT",
                     "renter_ratio_CAT","relative_chronic_dis_mort_CAT")
merged_data_for_table <- merged_data[, table_col_order]

 table<-merged_data_for_table%>%
  rowwise() %>%
  mutate(number_highs= sum(c_across(7:11) == "High priority", na.rm = TRUE),
         number_mediums= sum(c_across(7:11) == "Medium priority", na.rm = TRUE),
         temp_rank=(number_highs*2)+number_mediums
         )%>%
   ungroup()%>%
   arrange(desc(temp_rank))%>%
   select(-c(number_highs, number_mediums))%>%
   slice(1:15)
table
```

```
## # A tibble: 15 x 12
##    county    summed_total_cost pop12_sqmi med_age renter_ratio relative_chronic~
##    <chr>                 <dbl>      <dbl>   <dbl>        <dbl>             <dbl>
##  1 Amador           598970736.      63.3     48.2        0.253            0.102
##  2 Calaveras        131848234.      44.6     49.1        0.231            0.0875
##  3 Tuolumne         242129946       24.3     47.1        0.302            0.105
##  4 Inyo             169160700.       1.82    45.5        0.364            0.0737
##  5 Lake            1347450993.      49.1     45          0.342            0.102
##  6 Mariposa          17474756       12.6     49.2        0.321            0.0702
##  7 Nevada          6352716267.     103.      47.5        0.280            0.0984
##  8 Plumas            46955168        7.65    49.5        0.305            0.0750
##  9 Siskiyou        1558949981.       7.12    46.8        0.353            0.110
## 10 Tehama           610226591.      21.5     39.5        0.354            0.0896
## 11 Alpine                   0        1.54    46.4        0.282            0
## 12 Del Norte        615640530.      28.3     39          0.383            0.0776
## 13 El Dorado       4239088028.     102.      43.5        0.268            0.0776
## 14 Humboldt       17981394511.      38.1     37.3        0.450            0.0929
## 15 Modoc           1703663257        2.33    46          0.314            0.0687
## # ... with 6 more variables: med_age_CAT <fct>, summed_total_cost_CAT <fct>,
## #   pop12_sqmi_CAT <fct>, renter_ratio_CAT <chr>,
```

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

County	Previous spending on projects		Population density		Median age of population		% population that are renters		Chronic disease mortality burden	
Amador	598970736	High priority	63.288340	High priority	48.2	High priority	0.2530030	Low priority	0.1021536	High priority
Calaveras	131848234	High priority	44.582939	High priority	49.1	High priority	0.2311765	Low priority	0.0875314	High priority
Tuolumne	242129946	High priority	24.304973	High priority	47.1	High priority	0.3017241	Low priority	0.1051309	High priority
Inyo	169160700	High priority	1.819773	High priority	45.5	High priority	0.3637719	Low priority	0.0736661	Medium priority
Lake	1347450993	Medium priority	49.082334	High priority	45.0	High priority	0.3418713	Low priority	0.1024321	High priority
Mariposa	17474756	High priority	12.613887	High priority	49.2	High priority	0.3205512	Low priority	0.0701707	Medium priority
Nevada	6352716267	Medium priority	102.564339	High priority	47.5	High priority	0.2802273	Low priority	0.0983582	High priority
Plumas	46955168	High priority	7.653217	High priority	49.5	High priority	0.3054473	Low priority	0.0750500	Medium priority
Siskiyou	1558949981	Medium priority	7.120891	High priority	46.8	High priority	0.3525250	Low priority	0.1097566	High priority
Tehama	610226591	High priority	21.523312	High priority	39.5	Medium priority	0.3535995	Low priority	0.0895902	High priority
Alpine	0	High priority	1.543841	High priority	46.4	High priority	0.2816901	Low priority	0.0000000	Low priority
Del Norte	615640530	High priority	28.298164	High priority	39.0	Medium priority	0.3828606	Low priority	0.0776015	Medium priority
El Dorado	4239088028	Medium priority	102.156840	High priority	43.5	High priority	0.2681742	Low priority	0.0775971	Medium priority
Humboldt	17981394511	Medium priority	38.062105	High priority	37.3	Medium priority	0.4499474	Low priority	0.0928689	High priority
Modoc	1703663257	Medium priority	2.329272	High priority	46.0	High priority	0.3144685	Low priority	0.0687366	Medium priority

```
## #   relative_chronic_dis_mort_CAT <fct>, temp_rank <dbl>
```

```
table <- table %>% select(, c(1,2,8,3,9,4,7,5,10,6,11))

table_printed = kable(table,
        col.names = c("","","",
                      "", "",
                      "", "",
                       "", "",
                      "", ""
                    ),
       caption="Top 10 Counties ranked by need for oshpd projects.",
        booktabs=TRUE,
        align='lclclclclcl')%>%
   kable_styling(latex_options="scale_down")

add_header_above(table_printed, c("County" = 1,"Previous spending on projects" = 2,
                      "Population density" = 2,
                      "Median age of population" = 2,
                       "% population that are renters"= 2,
                      "Chronic disease mortality burden" = 2
                    ))
```