

# R Project Milestone 2

Moyra Rasheed, Courtney Coon, Jarett Maycott

2022-09-30

## Importing data

### demographic dataset

```
demo_data<-read.csv(  
  "https://raw.githubusercontent.com/PHW290/phw251_projectdata/main/ca_county_demographic.csv",  
  stringsAsFactors = FALSE)  
str(demo_data)
```

```
## 'data.frame':   58 obs. of  22 variables:  
## $ name       : chr  "Kern" "Kings" "Lake" "Lassen" ...  
## $ pop2012    : int   851089 155039 65253 35039 9904341 153025 255509 18455 88094 256841 ...  
## $ pop12_sqmi : num   104.28 111.43 49.08 7.42 2423.26 ...  
## $ white      : int   499766 83027 52033 25532 4936599 94456 201963 16103 67218 148381 ...  
## $ black      : int   48921 11014 1232 2834 856874 5629 6987 138 622 9926 ...  
## $ ameri_es   : int   12676 2562 2049 1234 72828 4136 1523 527 4277 3473 ...  
## $ asian      : int   34846 5620 724 356 1346865 2802 13761 204 1450 18836 ...  
## $ hawn_pi    : int   1252 271 108 165 26094 162 509 26 119 583 ...  
## $ hispanic   : int   413033 77866 11088 6117 4687889 80992 39069 1676 19505 140485 ...  
## $ other      : int   204314 42996 5455 3562 2140632 37380 16973 508 10185 62665 ...  
## $ mult_race  : int   37856 7492 3064 1212 438713 6300 10693 745 3970 11929 ...  
## $ males      : int   433108 86344 32469 22416 4839654 72682 124072 9269 43983 128737 ...  
## $ females    : int   406523 66638 32196 12479 4978951 78183 128337 8982 43858 127056 ...  
## $ med_age    : num    30.7 31.1 45 37 34.8 33.1 44.5 49.2 41.6 29.6 ...  
## $ households : int   254610 41233 26548 10058 3241204 43317 103210 7693 34945 75642 ...  
## $ families   : int   191739 31939 16255 6800 2194080 34093 62653 4948 21591 58767 ...  
## $ hse_units  : int   284367 43867 35492 12710 3445076 49140 111214 10188 40323 83698 ...  
## $ ave_fam_sz : num     3.61 3.59 2.94 2.98 3.58 3.63 2.94 2.77 3.02 3.74 ...  
## $ vacant     : int   29757 2634 8944 2652 203872 5823 8004 2495 5378 8056 ...  
## $ owner_occ  : int   152828 22329 17472 6590 1544749 27726 64637 5227 20601 41196 ...  
## $ renter_occ : int   101782 18904 9076 3468 1696455 15591 38573 2466 14344 34446 ...  
## $ county_fips: int    6103 6089 6106 6086 6073 6102 6066 6111 6100 6099 ...
```

```
#data is clean and ready to use
```

## mortality dataset

```
mort_data<-read.csv(
  "https://raw.githubusercontent.com/PHW290/phw251_projectdata/main/ca_county_mortality.csv",
  stringsAsFactors = FALSE, na.strings = "")

#remove last 2 columns b/c mostly NAs ("Annotation...")
mort_data<-mort_data[,1:8]
head(mort_data)
```

```
##   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
## 3 2014 Alameda      Occurrence      Age        1-4 years  ALL
## 4 2014 Alameda      Occurrence      Age        5-14 years  ALL
## 5 2014 Alameda      Occurrence      Age       15-24 years  ALL
## 6 2014 Alameda      Occurrence      Age       25-34 years  ALL
##           Cause_Desc Count
## 1 All causes (total)  9357
## 2 All causes (total)   105
## 3 All causes (total)    17
## 4 All causes (total)    17
## 5 All causes (total)   133
## 6 All causes (total)   175
```

```
#column names are all capitalized
colnames(mort_data) <- str_to_lower(colnames(mort_data))
head(mort_data)
```

```
##   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
## 3 2014 Alameda      Occurrence      Age        1-4 years  ALL
## 4 2014 Alameda      Occurrence      Age        5-14 years  ALL
## 5 2014 Alameda      Occurrence      Age       15-24 years  ALL
## 6 2014 Alameda      Occurrence      Age       25-34 years  ALL
##           cause_desc count
## 1 All causes (total)  9357
## 2 All causes (total)   105
## 3 All causes (total)    17
## 4 All causes (total)    17
## 5 All causes (total)   133
## 6 All causes (total)   175
```

```
#note "strata" and "strata_name" may have categories we want to pull out at some point
```

## HCAI Healthcare dataset

```
healthcare_data<-read.csv(  
  "https://raw.githubusercontent.com/PHW290/phw251_projectdata/main/hcai_healthcare_construction.csv",  
  stringsAsFactors = FALSE,na.strings = "")  
  
#remove last column ("Collection.of.Counties") b/c mostly NAs  
healthcare_data<-healthcare_data[,1:5]  
head(healthcare_data)
```

```
##           County Data.Generation.Date OSHPD.Project.Status  
## 1 01 - Alameda 2013-10-14T00:00:00      In Review  
## 2 01 - Alameda 2013-10-14T00:00:00 Pending Construction  
## 3 01 - Alameda 2013-10-14T00:00:00      In Construction  
## 4 01 - Alameda 2013-10-14T00:00:00      In Closure  
## 5 02 - Alpine 2013-10-14T00:00:00      In Review  
## 6 02 - Alpine 2013-10-14T00:00:00 Pending Construction  
## Total.Costs.of.OSHPD.Projects Number.of.OSHPD.Projects  
## 1                $50,890,315.00                44  
## 2                $840,242,543.36                125  
## 3                $994,245,713.95                181  
## 4                $65,337,613.88                 82  
## 5                  $0.00                      0  
## 6                  $0.00                      0
```

```
#column names are all capitalized  
colnames(healthcare_data) <- str_to_lower(colnames(healthcare_data))  
head(healthcare_data)
```

```
##           county data.generation.date oshpd.project.status  
## 1 01 - Alameda 2013-10-14T00:00:00      In Review  
## 2 01 - Alameda 2013-10-14T00:00:00 Pending Construction  
## 3 01 - Alameda 2013-10-14T00:00:00      In Construction  
## 4 01 - Alameda 2013-10-14T00:00:00      In Closure  
## 5 02 - Alpine 2013-10-14T00:00:00      In Review  
## 6 02 - Alpine 2013-10-14T00:00:00 Pending Construction  
## total.costs.of.oshpd.projects number.of.oshpd.projects  
## 1                $50,890,315.00                44  
## 2                $840,242,543.36                125  
## 3                $994,245,713.95                181  
## 4                $65,337,613.88                 82  
## 5                  $0.00                      0  
## 6                  $0.00                      0
```

## HCAI Healthcare dataset con't

```
#change "." to "_" for consistency with other datasets
names(healthcare_data) <- gsub(x = names(healthcare_data), pattern = "\\.",
                               replacement = "_")
head(healthcare_data)
```

```
##      county data_generation_date oshpd_project_status
## 1 01 - Alameda 2013-10-14T00:00:00      In Review
## 2 01 - Alameda 2013-10-14T00:00:00 Pending Construction
## 3 01 - Alameda 2013-10-14T00:00:00      In Construction
## 4 01 - Alameda 2013-10-14T00:00:00      In Closure
## 5 02 - Alpine 2013-10-14T00:00:00      In Review
## 6 02 - Alpine 2013-10-14T00:00:00 Pending Construction
## total_costs_of_oshpd_projects number_of_oshpd_projects
## 1                $50,890,315.00                44
## 2                $840,242,543.36                125
## 3                $994,245,713.95                181
## 4                $65,337,613.88                 82
## 5                  $0.00                      0
## 6                  $0.00                      0
```

```
#change county names to match other two datasets (remove numbers in front)
healthcare_data<-healthcare_data%>%
  mutate(county=substring(healthcare_data$county, 6))
head(healthcare_data)
```

```
##      county data_generation_date oshpd_project_status
## 1 Alameda 2013-10-14T00:00:00      In Review
## 2 Alameda 2013-10-14T00:00:00 Pending Construction
## 3 Alameda 2013-10-14T00:00:00      In Construction
## 4 Alameda 2013-10-14T00:00:00      In Closure
## 5 Alpine 2013-10-14T00:00:00      In Review
## 6 Alpine 2013-10-14T00:00:00 Pending Construction
## total_costs_of_oshpd_projects number_of_oshpd_projects
## 1                $50,890,315.00                44
## 2                $840,242,543.36                125
## 3                $994,245,713.95                181
## 4                $65,337,613.88                 82
## 5                  $0.00                      0
## 6                  $0.00                      0
```

## HCAI Healthcare dataset con't

```
#change money column ("Total.Costs...") to numeric
healthcare_data<-healthcare_data%>%
  mutate(total_costs_of_oshpd_projects = total_costs_of_oshpd_projects %>%
    str_remove_all("[$,]"))
healthcare_data$total_costs_of_oshpd_projects<-as.numeric(healthcare_data$total_costs_of_oshpd_projects)
head(healthcare_data)
```

```
##      county data_generation_date oshpd_project_status
## 1 Alameda 2013-10-14T00:00:00      In Review
## 2 Alameda 2013-10-14T00:00:00 Pending Construction
## 3 Alameda 2013-10-14T00:00:00      In Construction
## 4 Alameda 2013-10-14T00:00:00      In Closure
## 5 Alpine 2013-10-14T00:00:00      In Review
## 6 Alpine 2013-10-14T00:00:00 Pending Construction
##      total_costs_of_oshpd_projects number_of_oshpd_projects
## 1                      50890315                      44
## 2                      840242543                     125
## 3                      994245714                     181
## 4                      65337614                      82
## 5                          0                          0
## 6                          0                          0
```

```
#fix "data_generation_date" (remove empty time)
healthcare_data$data_generation_date<-as.Date(healthcare_data$data_generation_date)
head(healthcare_data)
```

```
##      county data_generation_date oshpd_project_status
## 1 Alameda      2013-10-14      In Review
## 2 Alameda      2013-10-14 Pending Construction
## 3 Alameda      2013-10-14      In Construction
## 4 Alameda      2013-10-14      In Closure
## 5 Alpine       2013-10-14      In Review
## 6 Alpine       2013-10-14 Pending Construction
##      total_costs_of_oshpd_projects number_of_oshpd_projects
## 1                      50890315                      44
## 2                      840242543                     125
## 3                      994245714                     181
## 4                      65337614                      82
## 5                          0                          0
## 6                          0                          0
```

## Description of dataset

**What is the data source? How does the dataset relate to the group problem statement and question?**

### **1. Demographic: Census Data**

Includes 58 observations with 22 variables. All are numeric or integers except for the first column which is a character with county names.

### **2. Mortality: CA open data portal, California Department of Public Health**

Includes 147,784 observations with 10 variables. most are characters except for year and count of mortality.

### **3. Healthcare: CA open data portal, Department of Healthcare Access and Information**

Includes 53,592 observations with 6 variables.

**How does the dataset relate to the group problem statement and question?**

Identify data types for 5+ data elements/columns/variables

- 1.
- 2.
- 3.
- 4.
- 5.

Utilize functions or resources in RStudio to determine the types of each data element (i.e. character, numeric, factor)

- 1.
- 2.
- 3.
- 4.
- 5.

Identify the desired type/format for each variable—will you need to convert any columns to numeric or another type?

- 1.
- 2.
- 3.
- 4.
- 5.

Provide a basic description of the 5+ data elements (Numeric: mean, median, range; Character: unique values/categories)

- 1.
- 2.
- 3.
- 4.
- 5.