Project_Milestone_2

Team 20

9/30/2021

#1: Data Description What is the data source? (1-2 sentences on where the data is coming from, dates included, etc.) The CalEnviroScreen3.0 data came from the California Environmental Protection Agency (CalEPA). It's publicly avaliable on the website of CaliforniaOffice of Environmental Health Hazard Assessment (OEHHA). It includes includes exposoure indicators (Ozone, PM2.5, diesel PM, drinking water contaminants, pesticide use, toxic release form facilities, traffic density), environemntal effect indicators(clean upsites, groundwater threats, hazard wastes generatosrs, impaired water bodies, solid waste sites), sensitive population indicators(asthma, cardiovascular diseases, low birth weight) and socioeconomic factor indicators(unemployment, housing, education, poverty, linguistic isolation).

Another dataset we use is a mortality dataset of California from 1989 to 2019. The dataset contains counts of deaths for California residents by ZIP Code based on information entered on death certificates.

How does the dataset relate to the group problem statement and question? Our group wants to look at the the relationship between air quality levels and Chronic Lower Respiratory Disease Mortality in California.The CalEnviroScreen3.0 data can provide us with PM2.5 air pollution data and the mortality data can provide us death from Chronic Lower Respiratory Disease.

#2: Import statement

Use appropriate import function and package based on the type of file. Utilize function arguments to control relevant components (i.e. change column types, column names, missing values, etc.). Document the import process

```
#air quality
air_qual<- read_csv("~/PHW251_Fall2021/AirQual_Project/air_qual.csv",
                    col_names = TRUE, col_types = NULL,  na = c("", "NA"))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   `California County` = col_character(),
##   ZIP = col_double(),
##   `Avg PM2.5 Per ZIP` = col_double()
## )
```

```
ca_air_qual <- rename(air_qual, county = "California County",
                      zip = "ZIP",
                      avg_PM2.5_per_zip = "Avg PM2.5 Per ZIP")
ca_air_qual[,3] <- round(ca_air_qual$avg_PM2.5_per_zip, 3)
print(ca_air_qual %>% head(3))
```

```
## # A tibble: 3 x 3
##   county    zip avg_PM2.5_per_zip
##   <chr>   <dbl>             <dbl>
## 1 Alameda 94501              8.70
## 2 Alameda 94502              8.70
## 3 Alameda 94536              8.94
```

```
dim(ca_air_qual)
```

```
## [1] 1377    3
```

```
#mort data
mort_by_zip <- read_csv("~/PHW251_Fall2021/AirQual_Project/mort_by_zip.csv",na = c("", "NA"))
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   X1 = col_double(),
##   Year = col_double(),
##   ZIP_Code = col_double(),
##   Geography_Type = col_character(),
##   Strata = col_character(),
##   Strata_Name = col_character(),
##   Cause = col_character(),
##   Cause_Desc = col_character(),
##   Count = col_double(),
##   Annotation_Code = col_double(),
##   Annotation_Desc = col_character()
## )
```

```
mort_by_zip <- select(mort_by_zip, "Year", "ZIP_Code", "Cause", "Cause_Desc", "Count")
mort_by_year <- filter(mort_by_zip, between(Year, 2009, 2018) )
```

```r
mort_by_cause <- filter(mort_by_year, Cause == "CLD")
mort_by_cause %>% head(2)
```

```
## # A tibble: 2 x 5
##    Year ZIP_Code Cause Cause_Desc                          Count
##   <dbl>    <dbl> <chr> <chr>                               <dbl>
## 1  2009    90001 CLD   Chronic lower respiratory diseases    NA
## 2  2009    90002 CLD   Chronic lower respiratory diseases    NA
```

#3: Identify data types for 5+ data elements/columns/variables.

Identify 5+ data elements required for your specified scenario. If <5 elements are required to complete the analysis, please choose additional variables of interest in the data set to explore in this milestone.Utilize functions or resources in RStudio to determine the types of each data element (i.e. character, numeric, factor) Identify the desired type/format for each variable—will you need to convert any columns to numeric or another type?

```
#var 1
class(ca_air_qual$county)
```

```
## [1] "character"
```

```
#var 2
class(ca_air_qual$zip)
```

```
## [1] "numeric"
```

```
ca_air_qual$zip <- as.character(ca_air_qual$zip)
class(ca_air_qual$zip)
```

```
## [1] "character"
```

```
# var 3
class(ca_air_qual$avg_PM2.5_per_zip)
```

```
## [1] "numeric"
```

```
#var 4
class(mort_by_cause$Year)
```

```
## [1] "numeric"
```

```
mort_by_cause$Year <- as.character(mort_by_cause$Year)
#var 5
class(mort_by_cause$Cause)
```

```
## [1] "character"
```

```
#var 6
mort_by_cause$ZIP_Code <- as.character(mort_by_cause$ZIP_Code)
class(mort_by_cause$ZIP_Code)
```

```
## [1] "character"
```

We needed to convert zipcodes into characters, instead of numeric values. We also needed to convert year into a character variable.

#4: Provide a basic description of the 5+ data elements.
Numeric: mean, median, range.
Character: unique values/categories.
Or any other descriptives that will be useful to the analysis.

```
#var 1
unique(ca_air_qual$county)
```

```
##  [1] "Alameda"         "Alpine"          "Amador"          "Butte"
##  [5] "Calaveras"       "Colusa"          "Contra Costa"    "Del Norte"
##  [9] "El Dorado"       "Fresno"          "Glenn"           "Humboldt"
## [13] "Imperial"        "Inyo"            "Kern"            "Kings"
## [17] "Lake"            "Lassen"          "Los Angeles"     "Madera"
## [21] "Marin"           "Mariposa"        "Mendocino"       "Merced"
## [25] "Modoc"           "Mono"            "Monterey"        "Napa"
## [29] "Nevada"          "Orange"          "Placer"          "Plumas"
## [33] "Riverside"       "Sacramento"      "San Benito"      "San Bernardino"
## [37] "San Diego"       "San Francisco"   "San Joaquin"     "San Luis Obispo"
## [41] "San Mateo"       "Santa Barbara"   "Santa Clara"     "Santa Cruz"
## [45] "Shasta"          "Sierra"          "Siskiyou"        "Solano"
## [49] "Sonoma"          "Stanislaus"      "Sutter"          "Tehama"
## [53] "Trinity"         "Tulare"          "Tuolumne"        "Ventura"
## [57] "Yolo"            "Yuba"
```

```
length(unique(ca_air_qual$county))
```

```
## [1] 58
```

```
#var 2
length(unique(ca_air_qual$zip))
```

```
## [1] 1355
```

```
#var 3
summary(ca_air_qual$avg_PM2.5_per_zip)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.651   7.860   9.536   9.781  11.840  19.309       7
```

```
#var 4
unique(mort_by_cause$Year)
```

```
##  [1] "2009" "2010" "2011" "2012" "2013" "2014" "2015" "2016" "2017" "2018"
```

```
#var 5
unique(mort_by_cause$Cause)
```

```
## [1] "CLD"
```

```
#var 6
length(unique(mort_by_cause$ZIP_Code))
```

```
## [1] 2664
```

```
#overall stats
str(ca_air_qual)
```

```
## spec_tbl_df [1,377 x 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ county           : chr [1:1377] "Alameda" "Alameda" "Alameda" "Alameda" ...
##  $ zip              : chr [1:1377] "94501" "94502" "94536" "94538" ...
```

```
##  $ avg_PM2.5_per_zip: num [1:1377] 8.7 8.7 8.94 9.44 9.46 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `California County` = col_character(),
##   ..   ZIP = col_double(),
##   ..   `Avg PM2.5 Per ZIP` = col_double()
##   .. )
```

```
str(mort_by_cause)
```

```
## tibble [26,640 x 5] (S3: tbl_df/tbl/data.frame)
##  $ Year      : chr [1:26640] "2009" "2009" "2009" "2009" ...
##  $ ZIP_Code  : chr [1:26640] "90001" "90002" "90003" "90004" ...
##  $ Cause     : chr [1:26640] "CLD" "CLD" "CLD" "CLD" ...
##  $ Cause_Desc: chr [1:26640] "Chronic lower respiratory diseases" "Chronic lower respiratory diseases
##  $ Count     : num [1:26640] NA NA 18 NA NA NA NA 15 0 NA ...
```