CDC

2024-04-10

Import Data

```
setwd("D:\\Northeastern\\Spring 2024\\DS5110\\DS5110---Heart-Disease-Analysis\\")

# CDC data
stoke_2016_2018 <- read.csv("cdc\\Heart_Disease_Mortality_Data_Among_US_Adults__35___by_State_Territory
stoke_2019_2021 <- read.csv("cdc\\Stroke_Mortality_Data_Among_US_Adults__35___by_State_Territory_and_Cou
```

Clean data

```
str(stoke_2016_2018)
```

```
## 'data.frame':    59094 obs. of  20 variables:
##  $ Year                       : int  2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 ...
##  $ LocationAbbr               : chr  "AK" "AK" "AK" "AK" ...
##  $ LocationDesc               : chr  "Aleutians East" "Aleutians West" "Anchorage" "Bethel" ...
##  $ GeographicLevel            : chr  "County" "County" "County" "County" ...
##  $ DataSource                 : chr  "NVSS" "NVSS" "NVSS" "NVSS" ...
##  $ Class                      : chr  "Cardiovascular Diseases" "Cardiovascular Diseases" "Cardiovascul
##  $ Topic                      : chr  "Heart Disease Mortality" "Heart Disease Mortality" "Heart Disea
##  $ Data_Value                 : num  173 172 243 337 NA ...
##  $ Data_Value_Unit            : chr  "per 100,000 population" "per 100,000 population" "per 100,000 po
##  $ Data_Value_Type            : chr  "Age-adjusted, Spatially Smoothed, 3-year Average Rate" "Age-adju
##  $ Data_Value_Footnote_Symbol: chr  "" "" "" "" ...
##  $ Data_Value_Footnote        : chr  "" "" "" "" ...
##  $ StratificationCategory1    : chr  "Gender" "Gender" "Gender" "Gender" ...
##  $ Stratification1            : chr  "Overall" "Overall" "Overall" "Overall" ...
##  $ StratificationCategory2    : chr  "Race/Ethnicity" "Race/Ethnicity" "Race/Ethnicity" "Race/Ethnici
##  $ Stratification2            : chr  "Overall" "Overall" "Overall" "Overall" ...
##  $ TopicID                    : chr  "T2" "T2" "T2" "T2" ...
##  $ LocationID                 : int  2013 2016 2020 2050 2060 2068 2070 2090 2100 2105 ...
##  $ Y_lat                      : num  55.4 53.6 61.2 60.9 58.8 ...
##  $ X_lon                      : num  -162 -167 -149 -160 -157 ...
```

```
str(stoke_2019_2021)
```

```
## 'data.frame':    78792 obs. of  21 variables:
##  $ Year                       : int  2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
##  $ LocationAbbr               : chr  "AK" "CT" "DE" "FL" ...
##  $ LocationDesc               : chr  "Nome" "Tolland County" "Delaware" "Washington County" ...
##  $ GeographicLevel            : chr  "County" "County" "State" "County" ...
##  $ DataSource                 : chr  "NVSS" "NVSS" "NVSS" "NVSS" ...
##  $ Class                      : chr  "Cardiovascular Diseases" "Cardiovascular Diseases" "Cardiovascul
##  $ Topic                      : chr  "Stroke Mortality" "Stroke Mortality" "Stroke Mortality" "Stroke
##  $ Data_Value                 : num  110.7 63.4 67.7 NA 69.5 ...
```

```
##  $ Data_Value_Unit           : chr  "per 100,000 population" "per 100,000 population" "per 100,000 po
##  $ Data_Value_Type           : chr  "Age-adjusted, Spatially Smoothed, 3-year Average Rate" "Age-adju
##  $ Data_Value_Footnote_Symbol: chr  "" "" "" "~" ...
##  $ Data_Value_Footnote       : chr  "" "" "" "Insufficient Data" ...
##  $ StratificationCategory1   : chr  "Gender" "Gender" "Gender" "Gender" ...
##  $ Stratification1           : chr  "Male" "Female" "Overall" "Female" ...
##  $ StratificationCategory2   : chr  "Race/Ethnicity" "Race/Ethnicity" "Race/Ethnicity" "Race/Ethnici
##  $ Stratification2           : chr  "Overall" "Asian" "Asian" "Hispanic" ...
##  $ TopicID                   : chr  "T6" "T6" "T6" "T6" ...
##  $ LocationID                : int  2180 9013 10 12133 17201 22107 25023 29217 29 38105 ...
##  $ Y_lat                     : num  64.9 41.9 39 30.6 42.3 ...
##  $ X_lon                     : num  -163.9 -72.3 -75.5 -85.7 -89.2 ...
##  $ Georeference              : chr  "POINT (-163.9462296 64.903977039)" "POINT (-72.337294 41.852989
```

```r
# drop Georeference column in stoke_2019_2021
drops <- c("Georeference")
stoke_2019_2021 <- stoke_2019_2021[ , !(names(stoke_2019_2021) %in% drops)]

cdc <- rbind(stoke_2016_2018, stoke_2019_2021)
cdc <- data.frame(cdc)

str(cdc)
```

```
## 'data.frame':    137886 obs. of  20 variables:
##  $ Year                      : int  2017 2017 2017 2017 2017 2017 2017 2017 2017 2017 ...
##  $ LocationAbbr              : chr  "AK" "AK" "AK" "AK" ...
##  $ LocationDesc              : chr  "Aleutians East" "Aleutians West" "Anchorage" "Bethel" ...
##  $ GeographicLevel           : chr  "County" "County" "County" "County" ...
##  $ DataSource                : chr  "NVSS" "NVSS" "NVSS" "NVSS" ...
##  $ Class                     : chr  "Cardiovascular Diseases" "Cardiovascular Diseases" "Cardiovascul
##  $ Topic                     : chr  "Heart Disease Mortality" "Heart Disease Mortality" "Heart Disea
##  $ Data_Value                : num  173 172 243 337 NA ...
##  $ Data_Value_Unit           : chr  "per 100,000 population" "per 100,000 population" "per 100,000 po
##  $ Data_Value_Type           : chr  "Age-adjusted, Spatially Smoothed, 3-year Average Rate" "Age-adju
##  $ Data_Value_Footnote_Symbol: chr  "" "" "" "" ...
##  $ Data_Value_Footnote       : chr  "" "" "" "" ...
##  $ StratificationCategory1   : chr  "Gender" "Gender" "Gender" "Gender" ...
##  $ Stratification1           : chr  "Overall" "Overall" "Overall" "Overall" ...
##  $ StratificationCategory2   : chr  "Race/Ethnicity" "Race/Ethnicity" "Race/Ethnicity" "Race/Ethnici
##  $ Stratification2           : chr  "Overall" "Overall" "Overall" "Overall" ...
##  $ TopicID                   : chr  "T2" "T2" "T2" "T2" ...
##  $ LocationID                : int  2013 2016 2020 2050 2060 2068 2070 2090 2100 2105 ...
##  $ Y_lat                     : num  55.4 53.6 61.2 60.9 58.8 ...
##  $ X_lon                     : num  -162 -167 -149 -160 -157 ...
```

```r
# drop unnecessary columns
drops <- c("Year", "X_lon", "Y_lat", "Class", "DataSource",
           "Data_Value_Footnote_Symbol", "Data_Value_Footnote",
           "StratificationCategory1", "StratificationCategory2",
           "Data_Value_Unit")
cdc <- cdc[ , !(names(cdc) %in% drops)]

# rename
colnames(cdc)[colnames(cdc) == "Data_Value"] <- "Data_Value_Per_100000_Population"
```

```r
for (col in names(cdc)) {
  if (is.numeric(cdc[[col]])) {
    mean_val <- round(mean(cdc[[col]], na.rm = TRUE), 2)
    cdc[[col]][is.na(cdc[[col]])] <- mean_val
  } else {
    mode_val <- names(sort(table(cdc[[col]]), decreasing = TRUE))[1]
    cdc[[col]][is.na(cdc[[col]])] <- mode_val
  }
}

str(cdc)
```

```
## 'data.frame':    137886 obs. of  10 variables:
##  $ LocationAbbr                    : chr  "AK" "AK" "AK" "AK" ...
##  $ LocationDesc                    : chr  "Aleutians East" "Aleutians West" "Anchorage" "Bethel" ...
##  $ GeographicLevel                 : chr  "County" "County" "County" "County" ...
##  $ Topic                           : chr  "Heart Disease Mortality" "Heart Disease Mortality" "Heart
##  $ Data_Value_Per_100000_Population: num  173 172 243 337 219 ...
##  $ Data_Value_Type                 : chr  "Age-adjusted, Spatially Smoothed, 3-year Average Rate" "Ag
##  $ Stratification1                 : chr  "Overall" "Overall" "Overall" "Overall" ...
##  $ Stratification2                 : chr  "Overall" "Overall" "Overall" "Overall" ...
##  $ TopicID                         : chr  "T2" "T2" "T2" "T2" ...
##  $ LocationID                      : num  2013 2016 2020 2050 2060 ...
```

Export to CSV

```r
folder_path <- "./cleaned-data/"  # Change this to your desired folder path

# Create the folder if it doesn't already exist
if (!dir.exists(folder_path)) {
  dir.create(folder_path)
}

file_path <- file.path(folder_path, "cleaned-cdc.csv")

# Export the dataframe to a CSV file
write.csv(cdc, file = file_path, row.names = FALSE)
```