Data Consolidation and Visualization

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor

Attaching package: 'zoo'


The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

AQI Data - Averaging and Consolidating

```
aqi1 <- read.csv("annual_aqi_by_county_2006.csv")
colnames(aqi1)
```

```
 [1] "State"                                "County"
 [3] "Year"                                 "Days.with.AQI"
 [5] "Good.Days"                            "Moderate.Days"
 [7] "Unhealthy.for.Sensitive.Groups.Days" "Unhealthy.Days"
 [9] "Very.Unhealthy.Days"                  "Hazardous.Days"
[11] "Max.AQI"                              "X90th.Percentile.AQI"
[13] "Median.AQI"                           "Days.CO"
[15] "Days.NO2"                             "Days.Ozone"
[17] "Days.PM2.5"                           "Days.PM10"
```

```
# aqi1
```

```
library(dplyr)
library(readr)
library(janitor)
```

```
Attaching package: 'janitor'
```

```
The following objects are masked from 'package:stats':

    chisq.test, fisher.test
```

```r
library(tidyr)


process_year_data <- function(year) {
  if (year == 2020) {
    return(NULL)
  }


  file_name <- paste0("annual_aqi_by_county_", year, ".csv")


  data <- read_csv(file_name, show_col_types = FALSE) %>%
    clean_names()


  required_columns <- c("max_aqi", "x90th_percentile_aqi", "median_aqi",
                        "days_with_aqi", "good_days",
                        "moderate_days",
                        "unhealthy_for_sensitive_groups_days",
                        "unhealthy_days","very_unhealthy_days",
                        "hazardous_days",
                        "days_co", "days_no2", "days_ozone",
                        "days_pm2_5", "days_pm10")


  missing_columns <- setdiff(required_columns, colnames(data))
  if (length(missing_columns) > 0) {
    warning("Missing columns in ", year, ": ",
            paste(missing_columns, collapse = ", "))
    return(NULL)
  }

  state_avg_aqi <- data %>%
    group_by(state) %>%
    summarise(
      avg_max_aqi = mean(max_aqi,
                         na.rm = TRUE),
      avg_x90th_percentile_aqi = mean(x90th_percentile_aqi,
                                      na.rm = TRUE),
      avg_median_aqi = mean(median_aqi,
                            na.rm = TRUE),
      avg_days_with_aqi = mean(days_with_aqi,
                               na.rm = TRUE),
      avg_good_days = mean(good_days,
```

```
                            na.rm = TRUE),
      avg_moderate_days = mean(moderate_days,
                               na.rm = TRUE),
      avg_unhealthy_for_sensitive_groups_days = mean
      (unhealthy_for_sensitive_groups_days, na.rm = TRUE),
      avg_unhealthy_days = mean(unhealthy_days,
                                na.rm = TRUE),
      avg_very_unhealthy_days = mean(very_unhealthy_days,
                                     na.rm = TRUE),
      avg_hazardous_days = mean(hazardous_days,
                                na.rm = TRUE),
      avg_days_co = mean(days_co,
                         na.rm = TRUE),
      avg_days_no2 = mean(days_no2,
                          na.rm = TRUE),
      avg_days_ozone = mean(days_ozone,
                            na.rm = TRUE),
      avg_days_pm2_5 = mean(days_pm2_5,
                            na.rm = TRUE),
      avg_days_pm10 = mean(days_pm10,
                           na.rm = TRUE)
    ) %>%
    mutate(year = year)

  return(state_avg_aqi)
}

# Process data for all years, excluding 2020
years <- setdiff(1999:2021, 2020)

all_years_data <- lapply(years, function(year) {
  data <- process_year_data(year)
  if (!is.null(data)) {
    return(data)
  }
}) %>%
  bind_rows()

# Create imputed 2020 data by averaging other years' data
impute_2020_data <- all_years_data %>%
  group_by(state) %>%
  summarise(
    avg_max_aqi = mean(avg_max_aqi,
                       na.rm = TRUE),
    avg_x90th_percentile_aqi = mean(avg_x90th_percentile_aqi,
                                    na.rm = TRUE),
    avg_median_aqi = mean(avg_median_aqi,
                          na.rm = TRUE),
```

```
    avg_days_with_aqi = mean(avg_days_with_aqi,
                             na.rm = TRUE),
    avg_good_days = mean(avg_good_days,
                         na.rm = TRUE),
    avg_moderate_days = mean(avg_moderate_days,
                             na.rm = TRUE),
    avg_unhealthy_for_sensitive_groups_days = mean
    (avg_unhealthy_for_sensitive_groups_days, na.rm = TRUE),
    avg_unhealthy_days = mean(avg_unhealthy_days,
                              na.rm = TRUE),
    avg_very_unhealthy_days = mean(avg_very_unhealthy_days,
                                   na.rm = TRUE),
    avg_hazardous_days = mean(avg_hazardous_days,
                              na.rm = TRUE),
    avg_days_co = mean(avg_days_co,
                       na.rm = TRUE),
    avg_days_no2 = mean(avg_days_no2,
                        na.rm = TRUE),
    avg_days_ozone = mean(avg_days_ozone,
                          na.rm = TRUE),
    avg_days_pm2_5 = mean(avg_days_pm2_5,
                          na.rm = TRUE),
    avg_days_pm10 = mean(avg_days_pm10,
                         na.rm = TRUE)
  ) %>%
  mutate(year = 2020)

# Combine the original data
aqi_final_data <- bind_rows(all_years_data, impute_2020_data)

write_csv(aqi_final_data, "state_avg_aqi_1999_2021_with_imputed_2020.csv")
```

```
ncol(aqi_final_data)
```

```
[1] 17
```

```
summary(aqi_final_data)
```

```
    state            avg_max_aqi      avg_x90th_percentile_aqi avg_median_aqi
 Length:1251        Min.   :  58.25   Min.   : 30.50           Min.   :15.25
 Class :character   1st Qu.: 101.03   1st Qu.: 56.13           1st Qu.:35.40
 Mode  :character   Median : 120.13   Median : 63.00           Median :40.95
                    Mean   : 133.11   Mean   : 66.93           Mean   :40.49
                    3rd Qu.: 146.17   3rd Qu.: 74.94           3rd Qu.:45.46
                    Max.   :1046.67   Max.   :141.00           Max.   :74.67
 avg_days_with_aqi avg_good_days   avg_moderate_days
 Min.   : 32.0     Min.   :  8.0   Min.   :  5.50
 1st Qu.:251.8     1st Qu.:153.4   1st Qu.: 62.44
 Median :288.2     Median :192.5   Median : 85.25
 Mean   :283.6     Mean   :189.5   Mean   : 85.80
```

```
 3rd Qu.:329.6     3rd Qu.:225.7   3rd Qu.:106.45
 Max.   :366.0     Max.   :356.2   Max.   :277.00
 avg_unhealthy_for_sensitive_groups_days avg_unhealthy_days
 Min.   : 0.000                          Min.   : 0.00000
 1st Qu.: 1.181                          1st Qu.: 0.04762
 Median : 3.688                          Median : 0.30450
 Mean   : 6.634                          Mean   : 1.42964
 3rd Qu.: 9.342                          3rd Qu.: 1.37798
 Max.   :58.667                          Max.   :26.33333
 avg_very_unhealthy_days avg_hazardous_days  avg_days_co
 Min.   : 0.00000        Min.   :0.00000    Min.   : 0.0000
 1st Qu.: 0.00000        1st Qu.:0.00000    1st Qu.: 0.0000
 Median : 0.00000        Median :0.00000    Median : 0.1333
 Mean   : 0.21024        Mean   :0.05039    Mean   : 3.9655
 3rd Qu.: 0.07596        3rd Qu.:0.00000    3rd Qu.: 1.9683
 Max.   :15.00000        Max.   :2.66667    Max.   :75.1333
  avg_days_no2      avg_days_ozone   avg_days_pm2_5   avg_days_pm10
 Min.   :  0.0000   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
 1st Qu.:  0.6085   1st Qu.: 95.27   1st Qu.: 69.16   1st Qu.:  0.50
 Median :  3.1333   Median :146.17   Median :104.27   Median :  8.00
 Mean   :  8.7656   Mean   :136.86   Mean   :115.55   Mean   : 18.49
 3rd Qu.: 10.0000   3rd Qu.:179.55   3rd Qu.:151.88   3rd Qu.: 23.52
 Max.   :143.0000   Max.   :291.00   Max.   :346.00   Max.   :173.50
      year
 Min.   :1999
 1st Qu.:2004
 Median :2010
 Mean   :2010
 3rd Qu.:2016
 Max.   :2021
```

```
str(aqi_final_data)
```

```
tibble [1,251 x 17] (S3: tbl_df/tbl/data.frame)
 $ state                                 : chr [1:1251] "Alabama" "Alaska" "Arizona" "Arkans
 $ avg_max_aqi                           : num [1:1251] 146 107 126 113 222 ...
 $ avg_x90th_percentile_aqi              : num [1:1251] 93.6 51.7 79.4 81.1 103.4 ...
 $ avg_median_aqi                        : num [1:1251] 54.4 23.7 48.8 55.4 51.6 ...
 $ avg_days_with_aqi                     : num [1:1251] 184 194 221 119 328 ...
 $ avg_good_days                         : num [1:1251] 69.8 167.7 104.2 69.3 177.1 ...
 $ avg_moderate_days                     : num [1:1251] 88.4 24.7 91.4 41.9 95.7 ...
 $ avg_unhealthy_for_sensitive_groups_days: num [1:1251] 20 1.5 23.33 6.39 36.91 ...
 $ avg_unhealthy_days                    : num [1:1251] 5.667 0.333 2.167 1.278 16.696 ...
 $ avg_very_unhealthy_days               : num [1:1251] 0.429 0 0 0 1.821 ...
 $ avg_hazardous_days                    : num [1:1251] 0 0 0 0 0.196 ...
 $ avg_days_co                           : num [1:1251] 2.238 49.333 0.25 0.833 3.339 ...
 $ avg_days_no2                          : num [1:1251] 0 0 12.67 2.11 45.95 ...
 $ avg_days_ozone                        : num [1:1251] 93.2 59.7 145.9 85.7 220.6 ...
 $ avg_days_pm2_5                        : num [1:1251] 71.5 50.2 37.8 30.2 45.8 ...
 $ avg_days_pm10                         : num [1:1251] 17.333 35 24.5 0.167 12.786 ...
```

```
 $ year                                  : num [1:1251] 1999 1999 1999 1999 1999 ...
```

Loading the "Cancer Incidence" Data

```
# Load the data
cancer_incidence <- read.csv("cancer_incidence.csv")

# Convert 'Count' and 'Population' to numeric
cancer_incidence$Count <- as.numeric(cancer_incidence$Count)
```

```
Warning: NAs introduced by coercion
```

```
cancer_incidence$Population <- as.numeric(cancer_incidence$Population)
```

```
Warning: NAs introduced by coercion
```

```
# Remove 'Crude.Rate' column
cancer_incidence <- cancer_incidence %>%
  select(-Crude.Rate)

# Aggregate data by State, Year (across both sexes)
cancer_aggregated <- cancer_incidence %>%
  group_by(States, Year) %>%
  summarise(
    Total_Count = sum(Count, na.rm = TRUE),
    Total_Population = sum(Population, na.rm = TRUE)
  )
```

```
`summarise()` has grouped output by 'States'. You can override using the
`.groups` argument.
```

```
# View the resulting aggregated data
head(cancer_aggregated)
```

```
# A tibble: 6 x 4
# Groups:   States [1]
  States   Year Total_Count Total_Population
  <chr>   <int>       <dbl>            <dbl>
1 Alabama  1999          41           194723
2 Alabama  2000          51           220789
3 Alabama  2001          79           442183
4 Alabama  2002          34           378534
5 Alabama  2003          53           474259
6 Alabama  2004          54           313205
```

Ensuring Data Integrity and Processing the Next Data Set

```
aqi_data <- aqi_final_data
can_in <- cancer_aggregated


aqi_data$year <- as.integer(as.character(aqi_data$year))
```

```r
can_in$Year <- as.integer(as.character(can_in$Year))


names(aqi_data)[names(aqi_data) == "state"] <- "States"
names(can_in)[names(can_in) == "States"] <- "States"


can_in_complete <- can_in %>%
  mutate(
    Total_Count = ifelse(is.na(Total_Count), 0, Total_Count),
    Total_Population = ifelse
    (is.na(Total_Population), 0, Total_Population)
  )

final_merged_data <- left_join(aqi_data, can_in_complete,
                               by = c("States", "year" = "Year"))

head(final_merged_data)
```

```
# A tibble: 6 x 19
  States     avg_max_aqi avg_x90th_percentile~1 avg_median_aqi avg_days_with_aqi
  <chr>            <dbl>                  <dbl>          <dbl>              <dbl>
1 Alabama           146.                   93.6           54.4               184.
2 Alaska            107                    51.7           23.7               194.
3 Arizona           126.                   79.4           48.8               221.
4 Arkansas          113.                   81.1           55.4               119.
5 California        222.                  103.            51.6               328.
6 Canada            133                    47             34                 188
# i abbreviated name: 1: avg_x90th_percentile_aqi
# i 14 more variables: avg_good_days <dbl>, avg_moderate_days <dbl>,
#   avg_unhealthy_for_sensitive_groups_days <dbl>, avg_unhealthy_days <dbl>,
#   avg_very_unhealthy_days <dbl>, avg_hazardous_days <dbl>, avg_days_co <dbl>,
#   avg_days_no2 <dbl>, avg_days_ozone <dbl>, avg_days_pm2_5 <dbl>,
#   avg_days_pm10 <dbl>, year <int>, Total_Count <dbl>, Total_Population <dbl>
```

```r
str(final_merged_data)
```

```
tibble [1,251 x 19] (S3: tbl_df/tbl/data.frame)
 $ States                                : chr [1:1251] "Alabama" "Alaska" "Arizona" "Arkans
 $ avg_max_aqi                           : num [1:1251] 146 107 126 113 222 ...
 $ avg_x90th_percentile_aqi              : num [1:1251] 93.6 51.7 79.4 81.1 103.4 ...
 $ avg_median_aqi                        : num [1:1251] 54.4 23.7 48.8 55.4 51.6 ...
 $ avg_days_with_aqi                     : num [1:1251] 184 194 221 119 328 ...
 $ avg_good_days                         : num [1:1251] 69.8 167.7 104.2 69.3 177.1 ...
 $ avg_moderate_days                     : num [1:1251] 88.4 24.7 91.4 41.9 95.7 ...
 $ avg_unhealthy_for_sensitive_groups_days: num [1:1251] 20 1.5 23.33 6.39 36.91 ...
 $ avg_unhealthy_days                    : num [1:1251] 5.667 0.333 2.167 1.278 16.696 ...
 $ avg_very_unhealthy_days               : num [1:1251] 0.429 0 0 0 1.821 ...
 $ avg_hazardous_days                    : num [1:1251] 0 0 0 0 0.196 ...
```

```
 $ avg_days_co                          : num [1:1251] 2.238 49.333 0.25 0.833 3.339 ...
 $ avg_days_no2                         : num [1:1251] 0 0 12.67 2.11 45.95 ...
 $ avg_days_ozone                       : num [1:1251] 93.2 59.7 145.9 85.7 220.6 ...
 $ avg_days_pm2_5                       : num [1:1251] 71.5 50.2 37.8 30.2 45.8 ...
 $ avg_days_pm10                        : num [1:1251] 17.333 35 24.5 0.167 12.786 ...
 $ year                                 : int [1:1251] 1999 1999 1999 1999 1999 1999 1999 1
 $ Total_Count                          : num [1:1251] 41 0 93 0 2028 ...
 $ Total_Population                     : num [1:1251] 194723 261961 475824 196611 3185892
```

```
write_csv(final_merged_data, "merged_aqi_cancer_incidence.csv")
```

Loading Environmental Hazard Data

```
narrowr <- read.csv("narrowresult.csv")
str(narrowr)
```

```
'data.frame':   273014 obs. of  23 variables:
 $ OrganizationIdentifier                          : chr  "AK-CHIN_WQX" "AK-CHIN_WQX" "AK-CHIN_
 $ OrganizationFormalName                          : chr  "Ak-Chin Indian Community (Tribal)" "
 $ ActivityIdentifier                              : chr  "AK-CHIN_WQX-SR:SD-23:2013-10-28" "AK
 $ ActivityStartDate                               : chr  "28/10/2013" "17/12/2013" "30/09/2013
 $ ResultDetectionConditionText                    : chr  "" "" "" "Not Reported" ...
 $ MethodSpecificationName                         : chr  "" "" "" "" ...
 $ CharacteristicName                              : chr  "Calcium" "Calcium" "Calcium" "Chlorc
 $ ResultSampleFractionText                        : chr  "Fixed" "Fixed" "Fixed" "" ...
 $ ResultMeasureValue                              : chr  "65.2" "56.3" "81.7" "" ...
 $ ResultMeasure.MeasureUnitCode                   : chr  "mg/L" "mg/L" "mg/L" "" ...
 $ ResultStatusIdentifier                          : chr  "Final" "Final" "Final" "Final" ...
 $ ResultValueTypeName                             : chr  "Actual" "Actual" "Actual" "Actual" "
 $ PrecisionValue                                  : num  NA NA NA NA NA NA NA NA NA NA ...
 $ DataQuality.BiasValue                           : logi  NA NA NA NA NA NA ...
 $ USGSPCode                                       : int  NA NA NA NA NA NA NA NA NA NA ...
 $ ResultDepthHeightMeasure.MeasureValue           : num  NA NA NA NA NA NA NA NA NA NA ...
 $ ResultDepthHeightMeasure.MeasureUnitCode        : chr  "" "" "" "" ...
 $ ResultDepthAltitudeReferencePointText           : chr  "" "" "" "" ...
 $ ResultSamplingPointName                         : chr  "" "" "" "" ...
 $ ResultAnalyticalMethod.MethodName               : chr  "Nitrate-Nitrite Nitrogen by Cd Reduc
 $ ResultAnalyticalMethod.MethodQualifierTypeName: chr  "" "" "" "" ...
 $ AnalysisStartDate                               : chr  "" "" "" "" ...
 $ AnalysisEndDate                                 : chr  "" "" "" "" ...
```

```
unique(narrowr$OrganizationFormalName)
```

```
  [1] "Ak-Chin Indian Community (Tribal)"
  [2] "ALABAMA DEPT. OF ENVIRONMENTAL MANAGEMENT - WATER QUALITY DATA"
  [3] "Animas River Stakeholders Group (Colorado) (Volunteer)"
  [4] "Arkansas Department of Environmental Quality"
  [5] "Big Valley Band of Pomo Indians of the Big Valley Rancheria, California (Tribal)"
  [6] "Boomsnub/Airco Superfund Site EPA Region 10"
  [7] "Bunker Hill Mining and Metallurgical Complex"
  [8] "Bureau of Reclamation"
  [9] "California Department Of Water Resources"
```

```
[10] "California Gulch (US EPA Region 8)"
[11] "California State Water Resources Control Board"
[12] "Captain Jack Mine (Colorado)"
[13] "CBS Operations Inc."
[14] "CDA TRUST"
[15] "CITY OF MARCO ISLAND"
[16] "Clear Creek Watershed Foundation (CCWF) (Volunteer)"
[17] "Coal Creek Watershed Coalition (Colorado)"
[18] "Collier County Coastal Zone Management Department (FL)"
[19] "Collier County Pollution Control (Florida)"
[20] "Colorado Dept. of Public Health & Environment-WQCD"
[21] "Colorado Division of Reclamation, Mining and Safety (DRMS) (Volunteer)"
[22] "Colorado Mountain College Natural Resource Management"
[23] "Colorado River Watch"
[24] "Connecticut Department Of Energy And Environmental Protection"
[25] "Cortina Rancheria (Kletsel Dehe Wintun Nation) (Tribal)"
[26] "Dade Environmental Resource Management (Florida)"
[27] "Division of Surface water (Ohio)"
[28] "EA Engineering, Science and Technology Inc."
[29] "EPA National Aquatic Resources Survey (NARS)"
[30] "EPA Region 10 Boomsnub Superfund Site Data 1987-2013"
[31] "EPA Region 10 Superfund Bunker Hill Mining and Metallurgical Complex"
[32] "EPA Region 4 Athens Lab (Georgia)"
[33] "FDEP GROUNDWATER MANAGEMENT SECTION"
[34] "FDEP TALLAHASSEE REGIONAL OPERATIONS CENTER"
[35] "FL Dept. of Environmental Protection"
[36] "FL Dept. of Environmental Protection, Northwest District"
[37] "Flandreau Santee Sioux Tribe (SD)"
[38] "Hopi Tribe of Arizona (Tribal)"
[39] "illinois epa"
[40] "Indiana STORET"
[41] "Jamestown S'Klallam Tribe (Tribal)"
[42] "Kickapoo Tribe of Indians of the Kickapoo Reservation in Kansas (Tribal)"
[43] "Lake County Water Resource Management"
[44] "Lake Fork Watershed Stakeholders (Colorado) (Volunteer)"
[45] "Massachusetts Department of Environmental Protection (MassDEP)"
[46] "Maul Foster and Alongi, Inc."
[47] "MBMG_WQX - Montana Bureau of Mines and Geology"
[48] "Midnite Mine Environmental Data"
[49] "Minnesota Pollution Control Agency - Ambient Surface Water"
[50] "Missouri Dept. of Natural Resources"
[51] "Montana DEQ WQPB"
[52] "Montana PPL Corporation"
[53] "Montana Volunteer Water Quality Monitoring"
[54] "Montana Watershed"
[55] "Morongo Band of Mission Indians (Tribal)"
[56] "Muckleshoot Indian Tribe (Tribal)"
[57] "National Park Service Water Resources Division"
[58] "Navajo Nation, Arizona, New Mexico & Utah (Tribal)"
```

```
 [59] "Nevada Division of Environmental Protection"
 [60] "New York State Dec Division Of Water"
 [61] "NM Environmental Dept./SWQB"
 [62] "North Dakota Department Of Environmental Quality"
 [63] "OCC - Otter Creek Coal"
 [64] "Oneida Nation"
 [65] "P4 Production LLC, Soda Springs Plant, Idaho"
 [66] "Palermo Wellfield Superfund Site by Geoengineers Inc. (Volunteer)*"
 [67] "Perry Co. Soil and Water District"
 [68] "Pueblo of Sandia Water Quality Program (New Mexico)"
 [69] "Red Lake DNR"
 [70] "Region 8 Superfund:  Standard Mine"
 [71] "Rhode Island"
 [72] "Salt Chuck Mine, State of Alaska"
 [73] "San Miguel Watershed Coalition (Volunteer)*"
 [74] "Santee Sioux Nation of Nebraska (Tribal)"
 [75] "Schuylkill Action Network (Pennsylvania)"
 [76] "Seminole Tribe of Florida (Tribal)"
 [77] "Shoalwater Bay Indian Tribe of the Shoalwater Bay Indian Reservation (Tribal)"
 [78] "Skagit County"
 [79] "Snoqualmie Indian Tribe (Tribal)"
 [80] "South Carolina Department of Environmental Services"
 [81] "Southwest Florida Water Management District"
 [82] "Spokane Tribe of the Spokane Reservation (Tribal)"
 [83] "State of Oregon Dept. of Environmental Quality"
 [84] "State of Wyoming Department of Environmental Quality Watershed Program"
 [85] "Suwannee River Water Management District"
 [86] "Table Mountain Rancheria of California (Tribal)"
 [87] "Tacoma-Pierce County Health Department (Washington)"
 [88] "TDEC Division of Water Resources"
 [89] "TerraGraphics Environmental Engineering, Inc."
 [90] "Texas Commission on Environmental Quality"
 [91] "Twenty-Nine Palms Tribal EPA"
 [92] "UD Citizen Monitoring Program"
 [93] "Uncompahgre Watershed Partnership (Volunteer)*"
 [94] "USEPA Region 9"
 [95] "USGS Florida Water Science Center"
 [96] "USGS Kansas Water Science Center"
 [97] "USGS Montana Water Science Center"
 [98] "USGS New Mexico Water Science Center"
 [99] "USGS Oregon Water Science Center"
[100] "Utah Department Of Environmental Quality"
[101] "Ute Mountain Utes Tribe (Colorado)"
[102] "VIRGINIA DEPARTMENT OF ENVIRONMENTAL QUALITY"
[103] "West Virginia Department of Environmental Protection Watershed Improvement Branch"
[104] "West Virginia Department of Environmental Protection-Division of Water & Waste Manager
[105] "Wind River Environmental Quality Commission"
[106] "Wisconsin Department of Natural Resources"
[107] "WV Div of Environmental Protection, Office of Water Resource"
```

```r
# List of U.S. state names
states <- c(
  "Alabama", "Alaska", "Arizona", "Arkansas", "California", "Colorado",
  "Connecticut", "Delaware", "Florida", "Georgia", "Hawaii", "Idaho",
  "Illinois", "Indiana","Iowa", "Kansas", "Kentucky", "Louisiana", "Maine",
  "Maryland", "Massachusetts", "Michigan", "Minnesota", "Mississippi",
  "Missouri", "Montana", "Nebraska","Nevada", "New Hampshire", "New Jersey",
  "New Mexico", "New York", "North Carolina","North Dakota", "Ohio",
  "Oklahoma", "Oregon", "Pennsylvania", "Rhode Island", "South Carolina",
  "South Dakota", "Tennessee", "Texas", "Utah", "Vermont","Virginia",
  "Washington", "West Virginia", "Wisconsin", "Wyoming"
)

# Regular expression pattern to match state names
state_pattern <- str_c(states, collapse = "|")


narrowr <- narrowr %>%
  mutate(
    # Handle blank or missing values first
    OrganizationFormalName = ifelse(is.na(OrganizationFormalName) |
                                      OrganizationFormalName == "", "Unknown",
                                    OrganizationFormalName),
    # Extract the state name if it exists in the organization name
    State = str_extract(OrganizationFormalName, state_pattern),
    # Replace organization name with state name if a match is found
    OrganizationFormalName = ifelse(!is.na(State), State,
                                    OrganizationFormalName)
  )
narrowr$state <- narrowr$OrganizationFormalName
# View the updated dataset
head(narrowr)
```

```
  OrganizationIdentifier            OrganizationFormalName
1            AK-CHIN_WQX Ak-Chin Indian Community (Tribal)
2            AK-CHIN_WQX Ak-Chin Indian Community (Tribal)
3            AK-CHIN_WQX Ak-Chin Indian Community (Tribal)
4            AK-CHIN_WQX Ak-Chin Indian Community (Tribal)
5            AK-CHIN_WQX Ak-Chin Indian Community (Tribal)
6            AK-CHIN_WQX Ak-Chin Indian Community (Tribal)
               ActivityIdentifier ActivityStartDate
1 AK-CHIN_WQX-SR:SD-23:2013-10-28        28/10/2013
2 AK-CHIN_WQX-SR:SD-23:2013-12-17        17/12/2013
3  AK-CHIN_WQX-SR:SD-23:2013-9-30        30/09/2013
4 AK-CHIN_WQX-SR:SD-23:2013-10-28        28/10/2013
5 AK-CHIN_WQX-SR:SD-23:2013-10-28        28/10/2013
6  AK-CHIN_WQX-SR:SD-23:2013-9-30        30/09/2013
  ResultDetectionConditionText MethodSpecificationName CharacteristicName
1                                                                 Calcium
```

```
2                                                                  Calcium
3                                                                  Calcium
4                Not Reported                                Chlorophyll a
5                                                              Potassium
6                                                                Sodium
  ResultSampleFractionText ResultMeasureValue ResultMeasure.MeasureUnitCode
1                    Fixed              65.2                         mg/L
2                    Fixed              56.3                         mg/L
3                    Fixed              81.7                         mg/L
4
5                                        6.5                         mg/L
6                    Fixed               114                         mg/L
  ResultStatusIdentifier ResultValueTypeName PrecisionValue
1                  Final              Actual             NA
2                  Final              Actual             NA
3                  Final              Actual             NA
4                  Final              Actual             NA
5                  Final              Actual             NA
6                  Final              Actual             NA
  DataQuality.BiasValue USGSPCode ResultDepthHeightMeasure.MeasureValue
1                    NA        NA                                   NA
2                    NA        NA                                   NA
3                    NA        NA                                   NA
4                    NA        NA                                   NA
5                    NA        NA                                   NA
6                    NA        NA                                   NA
  ResultDepthHeightMeasure.MeasureUnitCode
1
2
3
4
5
6
  ResultDepthAltitudeReferencePointText ResultSamplingPointName
1
2
3
4
5
6
                                ResultAnalyticalMethod.MethodName
1                      Nitrate-Nitrite Nitrogen by Cd Reduction
2                      Nitrate-Nitrite Nitrogen by Cd Reduction
3                      Nitrate-Nitrite Nitrogen by Cd Reduction
4                          Nitrite Nitrogen by Spectophotometry
5                          Nitrite Nitrogen by Spectophotometry
6 DO NOT USE***4500 NH3 C ~ Ammonia in Water by Titrimetric Method
  ResultAnalyticalMethod.MethodQualifierTypeName AnalysisStartDate
1
```

```
2
3
4
5
6                              duplicate records
  AnalysisEndDate State                                 state
1                  <NA> Ak-Chin Indian Community (Tribal)
2                  <NA> Ak-Chin Indian Community (Tribal)
3                  <NA> Ak-Chin Indian Community (Tribal)
4                  <NA> Ak-Chin Indian Community (Tribal)
5                  <NA> Ak-Chin Indian Community (Tribal)
6                  <NA> Ak-Chin Indian Community (Tribal)
```

```
colnames(narrowr)
```

```
 [1] "OrganizationIdentifier"
 [2] "OrganizationFormalName"
 [3] "ActivityIdentifier"
 [4] "ActivityStartDate"
 [5] "ResultDetectionConditionText"
 [6] "MethodSpecificationName"
 [7] "CharacteristicName"
 [8] "ResultSampleFractionText"
 [9] "ResultMeasureValue"
[10] "ResultMeasure.MeasureUnitCode"
[11] "ResultStatusIdentifier"
[12] "ResultValueTypeName"
[13] "PrecisionValue"
[14] "DataQuality.BiasValue"
[15] "USGSPCode"
[16] "ResultDepthHeightMeasure.MeasureValue"
[17] "ResultDepthHeightMeasure.MeasureUnitCode"
[18] "ResultDepthAltitudeReferencePointText"
[19] "ResultSamplingPointName"
[20] "ResultAnalyticalMethod.MethodName"
[21] "ResultAnalyticalMethod.MethodQualifierTypeName"
[22] "AnalysisStartDate"
[23] "AnalysisEndDate"
[24] "State"
[25] "state"
```

```
str(narrowr$ActivityStartDate)
```

```
 chr [1:273014] "28/10/2013" "17/12/2013" "30/09/2013" "28/10/2013" ...
```

```
str(narrowr$AnalysisStartDate)
```

```
 chr [1:273014] "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" ...
```

```
str(narrowr$AnalysisEndDate)
```

```
 chr [1:273014] "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" ...
```

```r
library(lubridate)
library(dplyr)

processed_dataset <- narrowr %>%
  # Drop specified columns
  select(-c(
    OrganizationIdentifier, state, OrganizationFormalName,
    ResultDepthAltitudeReferencePointText,
    ResultSamplingPointName,
    ResultAnalyticalMethod.MethodName,
    ActivityIdentifier, USGSPCode,
    ResultAnalyticalMethod.MethodQualifierTypeName,
    ResultDetectionConditionText,
    MethodSpecificationName, ResultStatusIdentifier,
    ResultSampleFractionText
  )) %>%


  mutate(

    ActivityStartDate = ifelse(ActivityStartDate == "" |
                                 is.na(ActivityStartDate), NA,
                               ActivityStartDate),
    AnalysisStartDate = ifelse(AnalysisStartDate == "" |
                                 is.na(AnalysisStartDate), NA,
                               AnalysisStartDate),
    AnalysisEndDate = ifelse(AnalysisEndDate == "" |
                               is.na(AnalysisEndDate),
                             NA, AnalysisEndDate),

    # Parse the dates with flexible parsing for character data
    ActivityStartDate = parse_date_time
    (ActivityStartDate, orders = c("dmy", "mdy", "ymd")),
    AnalysisStartDate = parse_date_time
    (AnalysisStartDate, orders = c("dmy", "mdy", "ymd")),
    AnalysisEndDate = parse_date_time
    (AnalysisEndDate, orders = c("dmy", "mdy", "ymd"))
  ) %>%

  mutate(
    AnalysisYear = case_when(
      !is.na(AnalysisEndDate) ~ year(AnalysisEndDate),
      !is.na(AnalysisStartDate) ~ year(AnalysisStartDate),
      !is.na(ActivityStartDate) ~ year(ActivityStartDate),
      TRUE ~ NA_real_
    )
  ) %>%
```

```
  # Drop rows where AnalysisYear is NA
  filter(!is.na(AnalysisYear)) %>%

  # Drop original date columns
  select(-c(ActivityStartDate, AnalysisStartDate, AnalysisEndDate))
```

```
Warning: There was 1 warning in `mutate()`.
i In argument: `AnalysisEndDate = parse_date_time(AnalysisEndDate, orders =
  c("dmy", "mdy", "ymd"))`.
Caused by warning:
! All formats failed to parse. No formats found.
```

```
# View the processed dataset
str(processed_dataset)
```

```
'data.frame':   273014 obs. of  10 variables:
 $ CharacteristicName                     : chr  "Calcium" "Calcium" "Calcium" "Chlorophyll
 $ ResultMeasureValue                     : chr  "65.2" "56.3" "81.7" "" ...
 $ ResultMeasure.MeasureUnitCode          : chr  "mg/L" "mg/L" "mg/L" "" ...
 $ ResultValueTypeName                    : chr  "Actual" "Actual" "Actual" "Actual" ...
 $ PrecisionValue                         : num  NA NA NA NA NA NA NA NA NA NA ...
 $ DataQuality.BiasValue                  : logi  NA NA NA NA NA NA ...
 $ ResultDepthHeightMeasure.MeasureValue  : num  NA NA NA NA NA NA NA NA NA NA ...
 $ ResultDepthHeightMeasure.MeasureUnitCode: chr  "" "" "" "" ...
 $ State                                  : chr  NA NA NA NA ...
 $ AnalysisYear                           : num  2013 2013 2013 2013 2013 ...
```

```
narrowrfilt <- processed_dataset %>%
  filter(rowSums(is.na(.) | . == "") < (ncol(processed_dataset) / 2))

# View the filtered dataset
head(narrowrfilt)
```

```
  CharacteristicName ResultMeasureValue ResultMeasure.MeasureUnitCode
1            Calcium               91.5                         mg/L
2          Magnesium                6.6                         mg/L
3            Calcium                 37                         mg/L
4            Calcium               90.5                         mg/L
5            Calcium               60.9                         mg/L
6          Magnesium                6.4                         mg/L
  ResultValueTypeName PrecisionValue DataQuality.BiasValue
1              Actual             NA                    NA
2              Actual             NA                    NA
3              Actual             NA                    NA
4              Actual             NA                    NA
5              Actual             NA                    NA
6              Actual             NA                    NA
  ResultDepthHeightMeasure.MeasureValue
1                                    NA
2                                    NA
3                                    NA
```

```
4                                      NA
5                                      NA
6                                      NA
  ResultDepthHeightMeasure.MeasureUnitCode    State AnalysisYear
1                                          Colorado         2015
2                                          Colorado         2015
3                                          Colorado         2015
4                                          Colorado         2015
5                                          Colorado         2015
6                                          Colorado         2015
```

```
nrow(narrowrfilt)
```

```
[1] 247307
```

```
colnames(narrowrfilt)
```

```
 [1] "CharacteristicName"
 [2] "ResultMeasureValue"
 [3] "ResultMeasure.MeasureUnitCode"
 [4] "ResultValueTypeName"
 [5] "PrecisionValue"
 [6] "DataQuality.BiasValue"
 [7] "ResultDepthHeightMeasure.MeasureValue"
 [8] "ResultDepthHeightMeasure.MeasureUnitCode"
 [9] "State"
[10] "AnalysisYear"
```

```
colnames(final_merged_data)
```

```
 [1] "States"
 [2] "avg_max_aqi"
 [3] "avg_x90th_percentile_aqi"
 [4] "avg_median_aqi"
 [5] "avg_days_with_aqi"
 [6] "avg_good_days"
 [7] "avg_moderate_days"
 [8] "avg_unhealthy_for_sensitive_groups_days"
 [9] "avg_unhealthy_days"
[10] "avg_very_unhealthy_days"
[11] "avg_hazardous_days"
[12] "avg_days_co"
[13] "avg_days_no2"
[14] "avg_days_ozone"
[15] "avg_days_pm2_5"
[16] "avg_days_pm10"
[17] "year"
[18] "Total_Count"
[19] "Total_Population"
```

```
colnames(narrowrfilt)
```

```
 [1] "CharacteristicName"
```

```
 [2] "ResultMeasureValue"
 [3] "ResultMeasure.MeasureUnitCode"
 [4] "ResultValueTypeName"
 [5] "PrecisionValue"
 [6] "DataQuality.BiasValue"
 [7] "ResultDepthHeightMeasure.MeasureValue"
 [8] "ResultDepthHeightMeasure.MeasureUnitCode"
 [9] "State"
[10] "AnalysisYear"
```

```r
final_merged_data <- final_merged_data %>%

  left_join(narrowrfilt, by = c("States" = "State", "year" = "AnalysisYear")) %>%
  mutate(across(everything(), ~replace(., is.na(.), "")))

# View the final merged dataset
head(final_merged_data)
```

```
# A tibble: 6 x 27
  States     avg_max_aqi avg_x90th_percentile~1 avg_median_aqi avg_days_with_aqi
  <chr>      <chr>       <chr>                  <chr>          <chr>
1 Alabama    145.523809~ 93.5714285714286       54.3809523809~ 184.238095238095
2 Alaska     107         51.6666666666667       23.6666666666~ 194.166666666667
3 Arizona    125.583333~ 79.4166666666667       48.75          221.166666666667
4 Arkansas   112.944444~ 81.0555555555556       55.4444444444~ 118.944444444444
5 California 222.321428~ 103.428571428571       51.5535714285~ 328.428571428571
6 California 222.321428~ 103.428571428571       51.5535714285~ 328.428571428571
# i abbreviated name: 1: avg_x90th_percentile_aqi
# i 22 more variables: avg_good_days <chr>, avg_moderate_days <chr>,
#   avg_unhealthy_for_sensitive_groups_days <chr>, avg_unhealthy_days <chr>,
#   avg_very_unhealthy_days <chr>, avg_hazardous_days <chr>, avg_days_co <chr>,
#   avg_days_no2 <chr>, avg_days_ozone <chr>, avg_days_pm2_5 <chr>,
#   avg_days_pm10 <chr>, year <chr>, Total_Count <chr>, Total_Population <chr>,
#   CharacteristicName <chr>, ResultMeasureValue <chr>, ...
```

```r
colnames(final_merged_data)
```

```
 [1] "States"
 [2] "avg_max_aqi"
 [3] "avg_x90th_percentile_aqi"
 [4] "avg_median_aqi"
 [5] "avg_days_with_aqi"
 [6] "avg_good_days"
 [7] "avg_moderate_days"
 [8] "avg_unhealthy_for_sensitive_groups_days"
 [9] "avg_unhealthy_days"
[10] "avg_very_unhealthy_days"
[11] "avg_hazardous_days"
[12] "avg_days_co"
[13] "avg_days_no2"
```

```
[14] "avg_days_ozone"
[15] "avg_days_pm2_5"
[16] "avg_days_pm10"
[17] "year"
[18] "Total_Count"
[19] "Total_Population"
[20] "CharacteristicName"
[21] "ResultMeasureValue"
[22] "ResultMeasure.MeasureUnitCode"
[23] "ResultValueTypeName"
[24] "PrecisionValue"
[25] "DataQuality.BiasValue"
[26] "ResultDepthHeightMeasure.MeasureValue"
[27] "ResultDepthHeightMeasure.MeasureUnitCode"
```

```
write_csv(final_merged_data, "final_dataset_consolidated.csv")
```

```
str(final_merged_data)
```

```
tibble [235,057 x 27] (S3: tbl_df/tbl/data.frame)
 $ States                                 : chr [1:235057] "Alabama" "Alaska" "Arizona" "Ark
 $ avg_max_aqi                            : chr [1:235057] "145.52380952381" "107" "125.5833
 $ avg_x90th_percentile_aqi               : chr [1:235057] "93.5714285714286" "51.666666666
 $ avg_median_aqi                         : chr [1:235057] "54.3809523809524" "23.666666666
 $ avg_days_with_aqi                      : chr [1:235057] "184.238095238095" "194.16666666
 $ avg_good_days                          : chr [1:235057] "69.7619047619048" "167.66666666
 $ avg_moderate_days                      : chr [1:235057] "88.3809523809524" "24.666666666
 $ avg_unhealthy_for_sensitive_groups_days : chr [1:235057] "20" "1.5" "23.3333333333333" "6
 $ avg_unhealthy_days                     : chr [1:235057] "5.66666666666667" "0.3333333333
 $ avg_very_unhealthy_days                : chr [1:235057] "0.428571428571429" "0" "0" "0"
 $ avg_hazardous_days                     : chr [1:235057] "0" "0" "0" "0" ...
 $ avg_days_co                            : chr [1:235057] "2.23809523809524" "49.333333333
 $ avg_days_no2                           : chr [1:235057] "0" "0" "12.6666666666667" "2.111
 $ avg_days_ozone                         : chr [1:235057] "93.1904761904762" "59.666666666
 $ avg_days_pm2_5                         : chr [1:235057] "71.4761904761905" "50.166666666
 $ avg_days_pm10                          : chr [1:235057] "17.3333333333333" "35" "24.5" "0
 $ year                                   : chr [1:235057] "1999" "1999" "1999" "1999" ...
 $ Total_Count                            : chr [1:235057] "41" "0" "93" "0" ...
 $ Total_Population                       : chr [1:235057] "194723" "261961" "475824" "19661
 $ CharacteristicName                     : chr [1:235057] "" "" "" "" ...
 $ ResultMeasureValue                     : chr [1:235057] "" "" "" "" ...
 $ ResultMeasure.MeasureUnitCode          : chr [1:235057] "" "" "" "" ...
 $ ResultValueTypeName                    : chr [1:235057] "" "" "" "" ...
 $ PrecisionValue                         : chr [1:235057] "" "" "" "" ...
 $ DataQuality.BiasValue                  : chr [1:235057] "" "" "" "" ...
 $ ResultDepthHeightMeasure.MeasureValue  : chr [1:235057] "" "" "" "" ...
 $ ResultDepthHeightMeasure.MeasureUnitCode: chr [1:235057] "" "" "" "" ...
```