CSCI 3900C Project IV (75 points)

Write an R script to accomplish the tasks described below. Please put all code in a single script file, and comment your code appropriately. Submit your script to the Project IV D2L drop box.

Background

Fine particulate matter (PM2.5) is a pollutant present in the air. There is strong evidence that PM2.5 is harmful to human health. In the U. S., the Environmental Protection Agency (EPA) sets national air quality standards for fine PM and tracks the emissions of PM2.5 into the atmosphere. About every 3 years, the EPA releases its database on emissions of PM2.5. This database is known as the National Emissions Inventory (NEI). See the EPA National Emissions Inventory web site for more details.

For each year and for each type of PM source, the NEI records how many tons of PM2.5 were emitted from that source over the course of the entire year.

Getting Started

Use the same data files that were posted for Project III. As a reminder, extract Project3.zip and use the following 2 files:

summarySCC_PM25.rds	*This file contains the data you will need to analyze*
Source_Classification_Code.rds	*This file classifies the source of pollutants (by SCC)*

The data file contains PM2.5 emissions data for 1999, 2002, 2005, and 2008.

You will need both files for this project. The appendix at the end of this document explains the content and format of these files.

Overview

You will investigate solvent-related pollutants in Los Angeles and two neighboring counties.

Program Tasks

Read each data file with the readRDS() function (see Project III for details).

Construct each of the following data frames.

1. A data frame containing only ***rows*** for these 3 California counties: Los Angeles County (`fips` code "06037"), Orange County ("06059"), and San Bernadino County ("06071"). The only ***columns*** in the data frame should be fips, SCC, and Emissions.
2. From the data frame constructed in step 1, add two additional columns: One called 'County' with the name of the county (based on fips code), and one called 'Solvent'. The value of 'Solvent' should be YES for any pollutant source that involves solvents, and NO otherwise.
3. From the data frame constructed in step 2, create a data frame with all years ***except*** 1999, and with only the columns County, year, Emissions, and Solvent.
4. Using your results from step 3, create a data frame that shows the total amount of Emissions for each county, by year, for each type of pollutant (solvent and non-solvent).

   Note that if this step is performed correctly, the data frame will have at most 18 rows (3 counties x 3 years x 2 pollutant types).
5. Again using your results from step 3, create a data frame that shows the total amount of Emissions for each county and each pollutant type (for 2002, 2005, and 2008 ***combined***).

6. Again with results from step 3, create a data frame that shows the total amount of ***solvent-related*** Emissions by year for each pollutant type (for the 3 counties ***combined***).
7. Create a data frame with one row for each county showing the total amount of Emissions for that county ***in 2008 only***.
8. Create a data frame ***with a single row*** showing the total amount of Emissions in all three counties combined, for all three years combined.

Appendix

The format and content of the data files are described below.

`summarySCC_PM25.rds` – *the PM2.5 Emissions Data file*

The file contains a data frame with all of the PM2.5 emissions data for 1999, 2002, 2005, and 2008. For each year, the table contains number of tons of PM2.5 emitted from a specific type of source for the entire year. Here are the first few rows:

```
##      fips      SCC Pollutant  Emissions type  year
##   4 09001 10100401 PM25-PRI      15.714 POINT 1999
##   8 09001 10100404 PM25-PRI     234.178 POINT 1999
##  12 09001 10100501 PM25-PRI       0.128 POINT 1999
##  16 09001 10200401 PM25-PRI       2.036 POINT 1999
##  20 09001 10200504 PM25-PRI       0.388 POINT 1999
##  24 09001 10200602 PM25-PRI       1.490 POINT 1999
```

The variables (columns) in the data frame are

- fips A string containing a 5-digit code representing the county
- SCC A digit string representing the specific pollutant source
- Pollutant A string identifying the pollutant
- Emissions Amount of PM2.5 emitted (tons)
- type The type of the source (point, non-point, on-road, non-road)
- year The year the emissions were recorded

`Source_Classification_Code.rds` – *a description of each pollutant source*

The file contains a table with information about each pollutant source. It can be used to look up the SCC digit string in the data file above. For each SCC digit string, this file provides a detailed name for the pollutant source, plus multiple classifications. (You will need this information to complete Task #2.) Here is ***part*** of the first row in the table:

SCC	**Data.Category**	**Short.Name**	**EI.Sector**	**SCC.Level.One**	**SCC.Level.Two**
10100101	Point	Ext Comb /Electric Gen /Anthracite Coal /Pulverized Coal	Fuel Comb - Electric Generation - Coal	External Combustion Boilers	Electric Generation