Project 4

Tyler Jones

2022-11-22

# Background

Sickle cell disease (SCD) is a group of genetic disorders in which red blood cells contort into a sickle shape. Symptoms of SCD include anemia (caused by misshapen cells dying early) as well as pain, infection, and stroke (caused by sickle cells blocking blood flow). There is some evidence that exposure to air pollution aggravates vaso-occlusion and pain in SCD patients. A hematologist at Duke wants to study the associations between various air pollutants and SCD in Durham County, North Carolina. She provides data on health system utilization (i.e., emergency department visits, hospitalizations, etc.) for SCD and you obtain air pollution data from the EPA. The pollutant data are raw measurements taken from the air pollution monitor at the RDU airport. This exposure data is messy and will need to be processed.

# Project Components

You have three .csv files containing daily air quality measurements for 2018-2020. Each file includes the date and concentrations of CO, PM2.5, and O3. Create a function to process the data. Run the function on the data in each .csv file. Your function should:

* Rename the pollutant concentration columns with more appropriate variable names.
* Format the date variable correctly.
* Some days have multiple measurements for a single pollutant. For these days, calculate the daily average.
* Remove any duplicate rows.
* Some of the pollutants have measurement error issues. Loop through the pollutants and set any measured concentration that is negative to zero.
* Create a table of the minimum and maximum of each pollutant.

# Function

library(tidyverse)  
library(flextable)  
library(here)  
set\_flextable\_defaults(font.family = "cambria")  
  
get\_tbl <- function(file){  
  
# load data  
dat <- read\_csv(paste0("data/",file),   
 col\_types = cols(Date = col\_date(format = "%m/%d/%Y"))) # Format date  
  
clean\_dat <- dat |>   
 rename(max.CO.conc = `Daily Max 8-hour CO Concentration`, # change col names  
 mean.PM2.conc = `Daily Mean PM2.5 Concentration`,  
 max.O3.conc = `Daily Max 8-hour Ozone Concentration`) |>   
 group\_by(Date) |>   
 summarise(across(.cols = where(is.numeric), .fns = ~mean(.x, na.rm = T))) |> # average vals for same day   
 ungroup() |>   
 mutate(across(.cols = where(is.numeric), .fns = ~ifelse(.x < 0, 0 , .x))) # if val < 0, make 0  
   
  
  
table <- clean\_dat |>   
 pivot\_longer(!Date, names\_to = "pollutant", values\_to = "conc") |>   
 group\_by(pollutant) |>   
 summarise(Minimum = round(min(conc, na.rm = T), 3), Maximum = round(max(conc, na.rm = T),3)) |> # Get max/min  
 ungroup() |>   
 arrange(factor(pollutant, levels = c("max.CO.conc", "mean.PM2.conc", "max.O3.conc"))) |> # order them right  
 mutate(pollutant = c("CO (ppm)", "PM2.5 (ug/m3)", "O3 (ppm)")) |> # Clean names  
 flextable(cwidth = 1.2) |>   
 set\_caption(  
 as\_paragraph(  
 as\_chunk(paste0("Table ",which(files == file),": Daily Concentrations"),   
 props = fp\_text\_default(italic = F))),   
 word\_stylename = "Table Caption") |>   
 align\_text\_col(align = "left") |>   
 align\_nottext\_col(align = "right") |>   
 line\_spacing(space = 0.5, part = "all") |>   
 void(part = "head", j = 1)  
  
table  
  
}

# Apply Function

files <- list.files(here("data"))  
  
tbls <- lapply(files, get\_tbl)

# Tables

**2018 Summary**

tbls[[1]]

Table 1: Daily Concentrations

|  | Minimum | Maximum |
| --- | --- | --- |
| CO (ppm) | 0.100 | 1.000 |
| PM2.5 (ug/m3) | 0.000 | 20.400 |
| O3 (ppm) | 0.012 | 0.067 |

**2019 Summary**

tbls[[2]]

Table 2: Daily Concentrations

|  | Minimum | Maximum |
| --- | --- | --- |
| CO (ppm) | 0.100 | 1.300 |
| PM2.5 (ug/m3) | 1.400 | 21.300 |
| O3 (ppm) | 0.015 | 0.068 |

**2020 Summary**

tbls[[3]]

Table 3: Daily Concentrations

|  | Minimum | Maximum |
| --- | --- | --- |
| CO (ppm) | 0.10 | 1.000 |
| PM2.5 (ug/m3) | 0.30 | 31.200 |
| O3 (ppm) | 0.01 | 0.052 |