Project 2

General Instructions

- Create an R project connected to your GitHub
- Your GitHub repo should contain all your code, data, and an informative README file
- Project output should be written in RMarkdown and submitted as a .rmd file that knits to a Word file
- Provide well-documented commentary on your code throughout your project & generally follow good programming practices
- Ensure your final report looks professional, is designed well, etc.
- Carefully consider the design of your plots
- Hide your code in your final Markdown document
- Projects are due November 17, 2025 at 5pm EDT
- Upload a .zip file of your R project folder to Canvas

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. You will write a function to process your data and do some exploratory EDA using plots.

Example data:

```
## # A tibble: 479 x 7
##   Date       Daily Max 8-hour CO Conc~1 units_CO Daily Mean PM2.5 Con~2 units_PM
##   <chr>                           <dbl> <chr>                      <dbl> <chr>
## 1 01/01/2018                        0.2 ppm                          7.5 ug/m3 LC
## 2 01/02/2018                        0.7 ppm                          7.5 ug/m3 LC
## 3 01/02/2018                        0.7 ppm                         10.4 ug/m3 LC
## 4 01/03/2018                        0.9 ppm                         15.2 ug/m3 LC
## 5 01/04/2018                        0.2 ppm                          9.7 ug/m3 LC
## 6 01/05/2018                        0.2 ppm                          5.7 ug/m3 LC
## 7 01/05/2018                        0.2 ppm                          9.8 ug/m3 LC
## 8 01/06/2018                        0.2 ppm                          7.6 ug/m3 LC
## # i 471 more rows
## # i abbreviated names: 1: `Daily Max 8-hour CO Concentration`,
## #   2: `Daily Mean PM2.5 Concentration`
## # i 2 more variables: `Daily Max 8-hour Ozone Concentration` <dbl>,
## #   units_O3 <chr>
```

Project Components

Function

You have three .csv files containing daily air quality measurements for 2018-2020. Each file includes the date and concentrations of CO, $PM_{2.5}$, and O_3. The 2020 dataset only includes data through April. 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 (no spaces, short, informative).
- 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. (Can also use apply() or map() functions).

Run your function on the three datasets.

Plots

You will create two plots.

- The first plot will have month on the x-axis and CO and O3 concentration on the y-axis. The plot should display monthly averages, averaged over the 3 years. Include 95% confidence interval error bars around the points. Make CO and O3 different colors.
- The second plot will have month on the x-axis and PM2.5 concentration on the y-axis. Include year as different colors. Use whatever geom you think best displays the data. There is no "correct" geom, but some are better than others. Consider readability - what conveys any patterns in your data best?

For both plots, be thoughtful in your color choices. Consider point size and density. Include labels, titles, legends, etc. - anything needed to make it readable.