Exercise - Cleaning Lead Exposure Datasets

In this exercise you will clean a pair of messy datasets on childhood lead exposure from the US states of Rhode Island (RI) and Maryland (MD). Each dataset contains childhool blood lead levels (BLL) measured in micrograms per deciliter (µg/dL) across a range of years and parts of the state, along with the corresponding number of children tested. Historically, BLLs above 10 were considered "elevated." More recently this threshold has been lowered to 5. The proportion of children above a given threshold can be viewed as a measure of the extent of harmful childhool lead exposure. All the files referred to below can be downloaded from the data directory of my personal website: `https://ditraglia.com/data`. This is a challenging exercise, but it is *exactly* the kind of thing that you will likely encounter when working with data in real life, either as part of your MPhil thesis research, as a research assistant, or both. To complete this exercise, you may find it helpful to consult this article on workflows with the `readxl` package.

1. The file `lead-RI-2005-2015.xlsx` contains "raw" BLL data for the state of Rhode Island from 2005 to 2015. Begin by downloading the relevant Excel file. Then open it in your favorite spreadsheet application to make sure you understand how it is set up. Your task is to use `readxl` and `dplyr` to load and clean the dataset so that it matches my own cleaned version of the same data: `lead-RI-2005-2015-cleaned.csv`. To be clear you *must* clean the data in R, not in Excel. The Rhode Island dataset is fairly simple in that it is contained in a single Excel sheet. It is more complicated in that it is given in *wide* format, so you will need to convert it to a *long* format tibble. Again, consult my cleaned dataset to see what the finished product should look like.

2. The file `lead-MD-2005-2015.xlsx` contains "raw" BLL data for the state of Maryland from 2005 to 2015. Begin by downloading the relevant Excel file. Then open it in your favorite spreadsheet application to make sure you understand how it is set up. Your task is to use `readxl` and `dplyr` to load and clean the dataset so it maches my own cleaned version of the same data: `lead-MD-2005-2015-cleaned.csv`. To be clear you *must* clean the data in R, not in Excel. The Maryland dataset is already in long format, but is stored across many sheets. You will need to find a way of importing and all of these an binding them together. Again: consult my cleaned dataset to see what the finished product should look like.

3. Now that you have clean datasets for RI and MD, use `dplyr` and `ggplot2` to explore them. What trends or patterns do you notice in childhood BLLs in these states and their regions between 2005 and 2015? Discuss briefly.