Purpose

This exercise is about data organization, orchestration, and coding, including creating Docker images.

Aim of this exercise is to:

1. Evaluate your coding (e.g. data reorganizations)
2. Understand data orchestration capabilities (e.g. Docker)
3. Understand how you design a solution (overall thinking)

Exercise

Data

The data is daily COVID case data from the United States. The data is located at: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports_us which is a series of CSV files, one for each day. Information about the data is located at https://github.com/CSSEGISandData/COVID-19.

Instructions

Put code in a public code repository hosted on GitHub, or a private repository and add `muschellij2` as a collaborator with read access

Deliverables

Note: the goal is the solution. If any steps below pose an unreasonable challenge at any time, and you need to get to the end result in a different way due to time constraints, please communicate that.

GitHub Repository/Actions

1. Create a GitHub repository for this exercise.

2. Create a Docker image that can read in the data from the day before, and compile a report/print out the cases for the day before. If the data is not there, print out a diagnostic message. If you are using R, you can use the rocker images as a base https://github.com/rocker-org/rocker-versioned2
3. Set up GitHub Actions to build this Docker image

Docker Image

Using this Docker image:

1. filter rows that are only in the United States,
2. take the mean cases (`Confirmed` variable) and deaths (`Deaths`) by state, averaging over counties (`Admin2`). Print this out in the action
3. Append the results to a file from the previous days' results.
4. run a this pipeline on a schedule (daily) using GitHub Actions.

Discussion/writeup

Please provide a half/full page description (either separate or in a README) of:

1. the challenges in getting this up and running
2. improvements you'd make in this pipeline if more time were available or any issues with the solution and how you'd perform checks on it
3. additional cleaning you would consider performing on a data set like this.