Assignment 01 - Exploratory Data Analysis

September 10, 2021

Due Date

This assignment is due at the end of the day September 24th, 2021

Learning Goals

- Download, read, and get familiar with an external dataset.
- Step through the EDA "checklist" presented in class
- Practice making exploratory plots

Assignment Description

We will work with air pollution data from the U.S. Environmental Protection Agency (EPA). The EPA has a national monitoring network of air pollution sites that The primary question you will answer is whether daily concentrations of $PM_{2.5}$ (particulate matter air pollution with aerodynamic diameter less than 2.5 μm) have decreased in California over the last 15 years (from 2004 to 2019).

A primer on particulate matter air pollution can be found here.

Your assignment should be completed in R markdown.

Steps

1. Given the formulated question from the assignment description, you will now conduct EDA Checklist items 2-4. First, download 2004 and 2019 data for all sites in California from the EPA Air Quality Data website. Read in the data using data.table(). For each of the two datasets, check the dimensions, headers, footers, variable names and variable types. Check for any data issues, particularly in the key variable we are analyzing. Make sure you write up a summary of all of your findings.

2. Combine the two years of data into one data frame. Use the Date variable to create a new column for year, which will serve as an identifier. Change the names of the key variables so that they are easier to refer to in your code.

3. Create a basic map in leaflet() that shows the locations of the sites (make sure to use different colors for each year). Summarize the spatial distribution of the monitoring sites.

4. Check for any missing or implausible values of $PM_{2.5}$ in the combined dataset. Explore the proportions of each and provide a summary of any temporal patterns you see in these observations.

5. Explore the main question of interest at three different spatial levels. Create exploratory plots (e.g. boxplots, histograms, line plots) and summary statistics that best suit each level of data. Be sure to write up explanations of what you observe in these data.

    - state
    - county
    - site in Los Angeles

This homework has been adapted from the case study in Roger Peng's Exploratory Data Analysis with R

George Vega Yon, Kim Siegmund, Abigail Horn
vegayon@usc.edu