HW 1: Due Day 1 11:59 PM

Exploratory Data Analysis

Use Canvas to submit the HW. Both RMarkdown and the knitted html file is required

Task 1

Instead of O_3 as used in the tutorial, perform the analysis for NO_2 for 2017 data. For this you will need to download the data from the EPA website and clean it. You need to tell a story about this pollutant and its distribution (in various senses). Explore some problematic features of the data collection and discuss their impacts on your analyses. You may consider some of the following questions in directing your analyses.

- How many sensors are measuring NO_2?
- Of what percentage of the year are the NO_2 sensors active? Where are the sensors that are not active? Map them.
- How does the geographical distribution of NO_2 AQI differ from that of O_3
- Which cities (CBSA, not counties) are worst affected by NO_2?
- Is there a correlation between NO_2 and O_3? Do different correlations matter? (i.e. correlations among AQIs of NO_2 and O_3 at site level, vs correlation of days AQI>100 at CBSA level, vs correlation of days AQI>100 at CBSA level etc.)
- What does the scatterplot look like? What does facetting the scatter plot by state tell us (pick 5 or so states)?
- Link this with other data (temperature, population etc.)? Where to acquires these datasets? How to link them?

These are by no means, exhaustive. Feel free to engage with your interests and the interesting bits about the datasets.

Task 2

Download the motor vehicle traffic collisons data from NYC Open data portal. Answer the following questions

- Which locations have high incidences of traffic collisons?
- How are these high traffic collisons locations different at different times of the day?
- Visualise the correlation between home values in a block group and traffic collisions and tell a story.