Citadel Datathon Team 11: An exploration of the effects of human activity on water pollution

Citadel Datathon Team 11: An exploration of the effects of human activity on water pollution

Summary

We investigate

1. Exploratory Data analysis

1.1 Exposure to pollution

Let $C = \{0, 1\}$ be the states of contamination, where 1 in particular denotes over than mean levels of contamination, and 0 otherwise. We first identify which states are most polluted. Let $M_{county,state}$ be the number of people in a selected county and state, exposed to level 1 of contamination (above mean), and let $N_{count,state}$ be the total number of population exposed to different states of contamination. The below geomap illustrates the population exposed to above mean levels of water contamination from years 1999 - 2016.

Since the above geomap is not weighted, it is more insightful to consider the ratio

$$R_{county} = \frac{M_{county}}{N_{county}}, \text{ and } R_{state} = \frac{M_{county,\cdot}}{N_{county,\cdot}},$$

where the notation $\cdot$ subscript describes $M_{state,\cdot} = \sum_{state} M_{county,state}$.

1.2 Exposed to types of chemicals

Once we identify our most polluted states the natural questions concern the type of pollution and the sources. Here we generate 3 graphs that demonstrate considerable distribution variation across our polluted states, in this case Florida, California and Connecticut.

Forest fires release large amounts of arsenic into the environment, which may explain the high levels in California. Connecticut's population also experiences high levels of above mean exposure in their population of arsenic. Interestingly, Florida exhibits contrasting levels, with low levels of abnormally high arsenic and uranium, and high levels of Halo-Acetic acid and trihalomethane.

Figure

1.3 Exposure to industry

2. Modelling

3. Experimental Results

4. Discussion