Data

The original csv file for this activity was downloaded from https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Selected-public-health-in/iqnk-2tcu. Go to that link for a more complete description of the data.

This lab is loosely based on https://data.world/exercises/cluster-analysis-exercise-2.

Your task

Use cluster analysis to identify groups communities that have characteristically similar public health statistics.

Considerations

- **Data Preparation** I've already removed columns with missing values and categorical variables. You should consider issues of normalization or scaling in your analysis.
- **Hyperparameters** What impact does changing the algorithm's parameters have on your results? Why did you make your final choice of parameters?
- **Interpretation** Is it possible to explain what each cluster represents? Did you retain or prepare a set of features that enables a meaningful interpretation of the clusters? Do the compositions of the clusters seem to make sense?
- **Important Note** This is an open-ended assignment (as many or most real-life data science projects are).

Extra Credit

- **Explore other clustering algorithms**
- **Validation** How will you measure the validity of your clustering process? Which metrics will you use, and how will you apply them? How do the results of the different algorithms compare under your metrics?
- **Correlation** Use the correlation coefficient to explore relationships between the different public health statistics.

Deliverables

In your Jupyter notebook, I expect to see

- implementation of a clustering algorithm,
- some visualization (this might take some imagination to think of ways to visualize high-dimensional data),
- a brief discussion of the process and rationale for your choices, and
- a brief explanation (interpretation) of the clusters.