Milestone 1: Website and ETL

Arturo Altamirano

My project will aim to prove that the industrial presence, particularly that of the petrochemical industry, is a positive correlator to cancer in the southern region. There was a study published by the Tulane University Law Center which claims that there will be negative impacts from the petrochemical industry in so called 'Cancer Alley' surrounding the Mississippi river and it's many tributaries in the Louisiana interior. I hope to uphold these projections with concrete data acquired from public sources. I will compare these county level metrics to a set of coordinates for the most prominent petrochemical plants in the region to see if there is a correlation. I hope to eventually develop a model to predict the severity of cancer rates in any given area based on proximity to petrochemical plants, as well as other factors such as average income, sewage system condition, and public health initiatives.

I am planning to use the following two datasets, sourced from Kaggle: Water Quality and Sewage System Dataset, and Cancer Regression | Cancer Data of the United States of America, as well as a self-created list of industrial sites collected by myself from various sources.

Data has been loaded and sorted. It can be viewed on the requested GitHub.io site. There may be a need to either synthetically generate data, or supplement the pollution sampling dataset with another, more ample dataset.