Rochester Dataset

Ming DeMers

Abstract

How does one quantify a city? GDP might allude to wealth, but how does one indicate a city's prosperity? This project will focus on answering those questions for the mid-sized city of Rochester, NY, by collecting and creating a dataset for several indicators. Methods in data sourcing and textual analysis are expected. A successful project will have a robust dataset that spans Rochester's history, a preliminary exploratory data analysis, and possibly an interactive data visualization.

Data

Data is to be collected manually from several sources. Namely, the US Census Bureau, the Rochester Open Data Portal. Additional information will be found from the Open Data Network, scraping headlines from local newspapers, RPD crime data, and possible sets from non-profits.

Approach

This project aims to produce three deliverables. Foremost is a cleaned and organized dataset(s) for future analysis. This will the foundation of a future project that analyzes growth and decay of rust belt cities. Data will be processed in the R language. Accompanying will be a codebook and exploratory data analysis, also written in R. These are simply references and documentation to show the basic shape, features, and extent of the dataset. It will be written in R and rendered with the Quarto Markdown engine. Finally, an interactive visualization will be written in Javascript and rendered by the Super 8 browser engine. This will allow users to physically work with the dataset in a hands-on and visual manner. The primary view will be a choropleth of the city of Rochester. A user will be able to zoom in and out, look over features of the map, filter for different values, go to certain neighborhoods, wards, etc. Secondary views will be graphs that show trends of the city, such as GDP, population, and more.

Significance

This project is the beginning of a larger research project to understand, analyze, and model the decay and projected growth of cities in the Rust Belt. Often, in data science, creating and cleaning datasets is the most tedious yet most important step. Therefore, devoting the semester to this project will be paramount to success in later steps. A robust dataset means more features to train of off, more dimensions for the simulation, and more reliability in the results. The success of the greater research project could suggest the ability to predict the trends of a city, and determine what changes would be necessary to support city success and growth.

References

[DSPS MSA Guide]
An online asynch course that details how to create a custom community dataset and apply data analysis techniques.

[Rochester Open Data Portal]
Rochester's open data portal, containing some 33 datasets of use.

[Open Data Rochester Datasets]
The Open Data Network, which contains a handful of additional datasets of use.