Fall 2024 Progress Report

Nathan Yang

Data Collection and Preparation

This project is a continuation of a Research Independent Study during the Spring 2024 Semester with the goal of the project to model academic performance in school districts across the United States through various demographic and socioeconomic factors.

This semester, I first reviewed and expanded the list of data sources for the project, incorporating new datasets identified through research papers and reports. To manage these sources, I maintained an Excel spreadsheet tracking the datasets and key variables of interest.

Data Integration Strategy Update

In the spring semester, I focused on datasets aggregated by school district, but this approach yielded limited results due to the scarcity of district-level data in ACS datasets. I pivoted to matching datasets at the county level, leveraging ACS Data Profiles (DP) for more comprehensive data coverage. This change significantly increased the volume and variety of data available for analysis (246 to 1388 observations).

Data Cleaning and Matching

Using the `readr`, `readxl`, `dplyr`, and `haven` packages, I cleaned and prepared datasets for integration. I employed fuzzy matching techniques with `stringdist` to join datasets with non-exact school district names. Metrics used included string distance and Jaccard difference, with thresholds determined from distribution analysis. Additional filtering criteria were applied to improve match accuracy, such as ensuring names began with the same letter and ended with the same three characters.

This approach significantly improved match rates. For instance:

CCD Data Join: Matched 4576 records, a 3% increase from the 4441 matches achieved using exact name matching.

ACS Income Data Join: Matched 281 records, a 12% increase from the 250 matches achieved using exact name matching.

As a result of the fuzzy matching and cleaning process, I successfully created a large, integrated dataset that combines data from multiple sources with significantly improved record retention. This dataset includes more comprehensive coverage of school district and county-level attributes, which will support more robust analysis and model development.

Next Steps

The new integrated dataset is larger and more feature-rich than the previous version, though the shift from district-level to county-level aggregation may impact model accuracy. This will be explored further during the modeling phase.

As the data preparation, cleaning, and integration phase has concluded, I will transition to the modeling stage which entails developing advanced models to explore the relationships between academic performance and demographic, socioeconomic, and geographic factors.

The revised timeline for this project is as follows:

Month	Milestone
January	Begin Model Development
February	Refine Models and Thesis Draft
March	Finalize Models and Thesis Final Draft
April (first 2 weeks)	Thesis Presentation/Defense

Reflection

The transition from district-level to county-level aggregation has been a significant improvement, enabling a more comprehensive dataset for analysis. The fuzzy matching techniques have also proven effective in enhancing record retention and matching accuracy.

I wish that I had identified the Census DPs earlier in the project as they have been a very valuable resource for consistent data across counties. This would have allowed me to save a lot of time and effort in the data integration process.