Group: Ascending

Repo Link: https://github.com/macs30122-winter25/final-project-ascending.git

Please see out notebooks for more information on the progress. Scapped data is *data/schools_il.csv*

Winter 2025 – Final Project Progress Report 1

Research Questions

Our research aims to answer the following questions:

1. How does school performance in Illinois commuting zones relate to intergenerational mobility within those zones?
2. Does higher school performance, as indicated by GreatSchools.org ratings, facilitate upward mobility across generations, or do structural barriers limit this relationship?

Originally focused solely on Illinois, our project now examines school performance in commuting zones across Wisconsin, Illinois, Indiana, and Michigan. By broadening the geographic scope and sample size, we aim to enhance the generalizability of our findings.

Key Variables and Operationalization

- **Dependent Variable:** Intergenerational mobility, defined as the extent to which individuals improve their socioeconomic standing relative to their parents. We obtain this data from publicly available geospatial datasets at Equality of Opportunity Project, expanding on research by Alesina, Stantcheva, and Teso (2018).
- **Independent Variable:** School performance within commuting zones, aggregated from individual school ratings on GreatSchools.org. These ratings include factors such as test scores, racial/ethnic composition, and disability statistics. Our analysis categorizes school levels into Pre-K, Elementary, Middle, and High School to examine differential impacts on mobility.

Methodology

We will first visualize the relationship between intergenerational mobility and school ratings at the commuting zone level. Our approach includes:

1. **Correlation Analysis:** Examining associations between school performance and mobility rates.
2. **Machine Learning Techniques:**
    - **Classification Model:** Categorizing school parameters based on their impact on intergenerational mobility.
    - **Regression Analysis:** Quantifying the influence of school characteristics (e.g., teacher-student ratio, student demographics) on mobility outcomes.

Compared to our initial proposal, we have decided to incorporate official data (e.g., college graduation rates, high school completion rates, and school quality indicators) from state education departments as a standardized measure of school performance. This refinement enhances the robustness of our analysis by providing a more comprehensive assessment of educational quality.

Our **unit of analysis** starts at the school level, where we merge all relevant data (e.g., GreatSchools and state datasets). Next, we aggregate school-level data to the tract level. Finally, we use the State-County FIPS code to link tracts to commuting zones, which is the level at which intergenerational mobility data is analyzed.

Data Cleaning and Wrangling

1. Datasets downloaded from the four states' education departments need to be cleaned. There are many redundancies and unneeded columns in them. In addition, we need to standardize the datasets to ensure consistency across sources. Some datasets contain rows

with school district names, while others have rows for individual schools or classes. We plan to merge these education data sources with school performance data (e.g., school.il.csv) scraped from GreatSchools.org. This integration will provide a more comprehensive dataset for machine learning analysis.

2. We used a national zip code CSV file to identify all zip codes in our target states. Using these zip codes, we scraped school data from GreatSchools.org, allowing us to analyze school performance at a regional level.

Data Analysis and Visualization

Variables and Construction

- **Independent Variable:** School performance, collected via the GreatSchools API and aggregated at the commuting zone level to explore regional patterns in school quality and intergenerational mobility.
- **Dependent Variable:** Relative intergenerational mobility, mapped using data from Alesina, Stantcheva, and Teso (2018).

We aggregate school-level ratings to the commuting zone level to match them with intergenerational mobility data. Additionally, we obtain the latitudes and longitudes of each school from the GreatSchools API to align school locations with commuting zones.

Data Analysis Methods

1. Our **unit of analysis** starts at the school level, where we merge all available data from GreatSchools and the four states' education departments. Next, we aggregate school performance data at the tract level. Finally, using the State-County FIPS code, we link tracts to commuting zones, the level at which intergenerational mobility data is analyzed.

2. We built a converter to process geographic information at the tract level. Schools were assigned to respective tracts using geographic data via **geopandas**.

3. **Panel Data Analysis**

- **Linear Regression:** Linear Regression Analysis shows that education-related variables—such as college graduation rates, high school completion rates, and school quality indicators—are significantly associated with higher mobility outcomes, even after controlling for economic, demographic, and infrastructure factors.

  Sensitivity Analysis further confirms that among all predictors, education-related factors exhibit strong and consistent effects on intergenerational mobility. Regions with higher graduation rates and better education quality tend to experience greater mobility across generations.

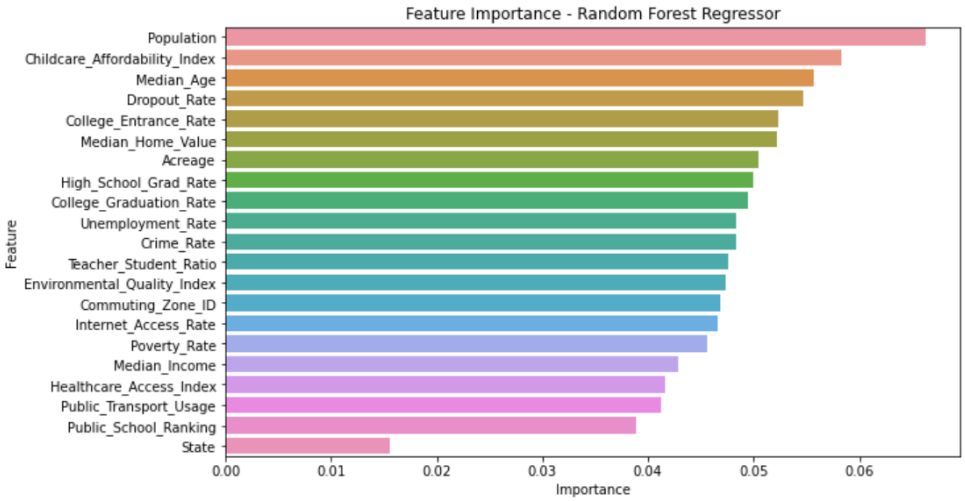
- **Random Forest Analysis:** Sensitivity analysis confirms that among all predictors, education-related factors exhibit strong and consistent effects on intergenerational mobility. Regions with higher graduation rates and better education quality tend to experience greater mobility across generations.

  Model Validation & Residual Analysis confirms that our findings are robust. After pruning the model to retain only the most impactful variables, residual diagnostics show that the model generalizes well, with residuals closely following a standard normal distribution. This suggests that education's impact is not an artifact of overfitting but a genuine and reliable effect.

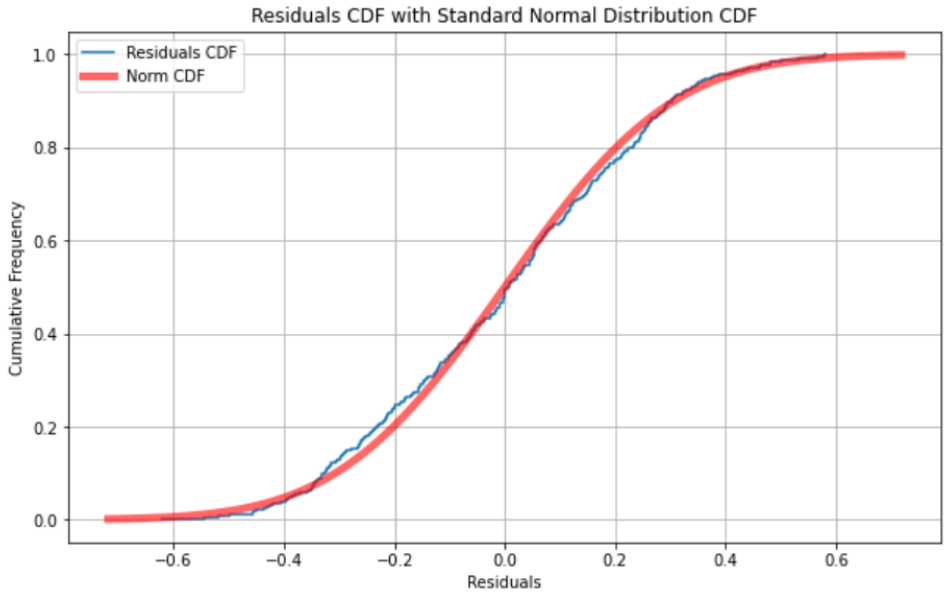
Expected Visualizations

- **Geographic heatmaps** displaying intergenerational mobility rates and school performance across commuting zones.

- **Regression plots** illustrating the relationship between school quality and mobility outcomes.



- **Random forest feature importance graphs** to highlight the most influential factors affecting mobility.



Team Responsibilities

1. **Jeanette:** Panel data analysis, regression modeling, causal pattern identification.
2. **Kunjian:** API maintenance, data collection and cleaning, GitHub repository management.

3. **Shirley:** Geographic data processing, data manipulation, map visualization.
4. **Carrie:** Collecting and cleaning official education datasets, presentation slides, README, final report.

Through our analyses, we expect to demonstrate that education plays a crucial role in shaping intergenerational mobility, reinforcing the importance of equitable access to quality schooling.