Final Memo

By Zhouxiang Sun

1. THE CIVIC PROBLEM ADDRESSED

Education does not exist in a vacuum. In Chicago, a student's academic trajectory is often inextricably linked to the socioeconomic fabric of their neighborhood. While the city collects vast amounts of data on school performance, these metrics are frequently viewed in isolation from the community context—poverty rates, income levels, and neighborhood safety.

The core civic problem addressed in this project is the systemic inequality in educational outcomes driven by external socioeconomic factors. Specifically, we investigate why college enrollment rates vary drastically across the city and whether "school quality" is merely a proxy for "neighborhood wealth." By integrating school-level data with community-level census indicators, this project aims to move beyond simple rankings to identify the root structural causes of educational disparity.

2. STEPS TAKEN & METHODOLOGY

To analyze this complex issue, I utilized R, a statistical programming language, to merge disparate datasets from the Chicago Data Portal. The analysis followed a rigorous data science workflow:

Step A: Data Integration & Cleaning I combined school profile data with socioeconomic indicators based on community areas. A critical step was handling missing data to ensure accurate modeling. The code logic involved performing a left-join operation between the schools dataset and the census income dataset, followed by removing incomplete records to prepare for machine learning.

Step B: Visualizing the Divide Using an interactive mapping tool (Leaflet), I overlaid school locations onto community wealth maps. This allowed for an immediate visual assessment of how high-performing schools cluster in affluent areas.

Step C: Advanced Predictive Modeling To determine which factors *actually* drive college enrollment, I moved beyond correlation to causation using Machine Learning models:

- Decision Tree: To visualize the decision-making path and identify the most critical thresholds.
- XGBoost: To validate findings and predict outcomes for schools with missing

data.

3. KEY FINDINGS

The analysis yielded three critical insights that challenge traditional assumptions about school performance:

Finding 1: Safety is the "Gateway" to Enrollment Contrary to the expectation that "Instruction Score" (teaching quality) is the top predictor, the Decision Tree analysis revealed that Safety Score is the primary determinant of college enrollment.

- Threshold: Schools with a safety score below 62 (out of 100) see a dramatic drop in college enrollment rates (averaging ~52%), regardless of instruction quality.
- Implication: Students cannot learn effectively if they do not feel safe.

Finding 2: The Wealth-Education Gradient The interactive map and correlation analysis confirm a strong spatial correlation. High college enrollment rates are almost exclusively concentrated in neighborhoods with low poverty and high middle-class populations (North Side/Downtown). Conversely, schools in high-poverty areas struggle to break past the 50% enrollment mark.

Finding 3: The Limits of Demographics While poverty is a strong predictor, the "Old-Age Rate" (percentage of seniors) showed a negligible linear relationship with instruction quality. This suggests that school performance is driven more by economic resources and safety than by the generational age structure of the neighborhood.

4. EQUITY AND FAIRNESS CONCERNS

From a data ethics perspective, two major equity concerns emerged during this project:

1. The "Data Void" in Vulnerable Communities Our analysis found that missing values for "College Enrollment Rate" were not random. They occurred disproportionately in high-poverty neighborhoods.

- Concern: By failing to collect or report data for these struggling schools, we render their challenges invisible. Policy decisions based on "available data" will inherently exclude the most vulnerable populations.

2. The Double Penalty of Safety The machine learning model shows that low safety scores place a "ceiling" on academic achievement.

- Concern: Students in low-income areas face a double penalty: they lack economic resources at home and attend schools where low safety scores

physically impede learning. Judging these schools solely on academic output without adjusting for the safety environment is structurally unfair.

5. POLICY IMPLICATIONS

Based on these findings, I propose the following recommendations for city leadership:

1. Prioritize "Safety-First" Intervention Zones Since the data indicates a safety score of 62 is a critical tipping point, the city should designate schools below this threshold as "Intervention Zones."

- Action: Allocate budget not just for textbooks, but for "Safe Passage" programs, mental health counselors, and infrastructure improvements in these specific zones to raise safety scores above 62.

2. Holistic Resource Allocation Funding formulas should act as a counterbalance to neighborhood poverty, not a reflection of it.

- Action: Use the economic index and our predicted missing enrollment data to direct supplemental grants to schools that are currently falling into the "data void."

3. Data Transparency Reform The city must address the missing data in high-poverty areas.

- Action: Implement mandatory reporting protocols for alternative schools and those in under-served areas to ensure they are represented in city-wide analytics. We cannot fix what we do not measure.