COLUMBIA UNIVERSITY

Impact of Subway Station Centrality on Manhattan Public High School Performance

Caroline Smyth [1] Kathy Xu [1] Sandra Zelen [1]

Introduction

The New York City Metropolitan Transit Authority subway system is the largest transportation agency in North America, serving more than 3 million riders daily. In Manhattan, only 23.4% of households own a car, making the borough highly dependent on the subway for transportation [1].

Research Topic

Understanding that robust transportation infrastructure greatly influences urban landscapes, our study explores the impact of the MTA on public high school performance throughout Manhattan. Previous research has studied how subway centrality shapes various socioeconomic outcomes, such as job mobility, racial inequities, and housing prices. However, there is a lack of literature on the relationship between subway centrality and education quality. This gap is worth exploring because education quality has been linked to long-term educational attainment and higher valuations of neighborhoods [5]. A more holistic understanding of how subway infrastructure affects these outcomes can inform school choice, and possibly contribute to urban planning decisions like deciding locations for new schools or where to extend subway lines.

Data Processing

We drew from two major sources for our data collection:

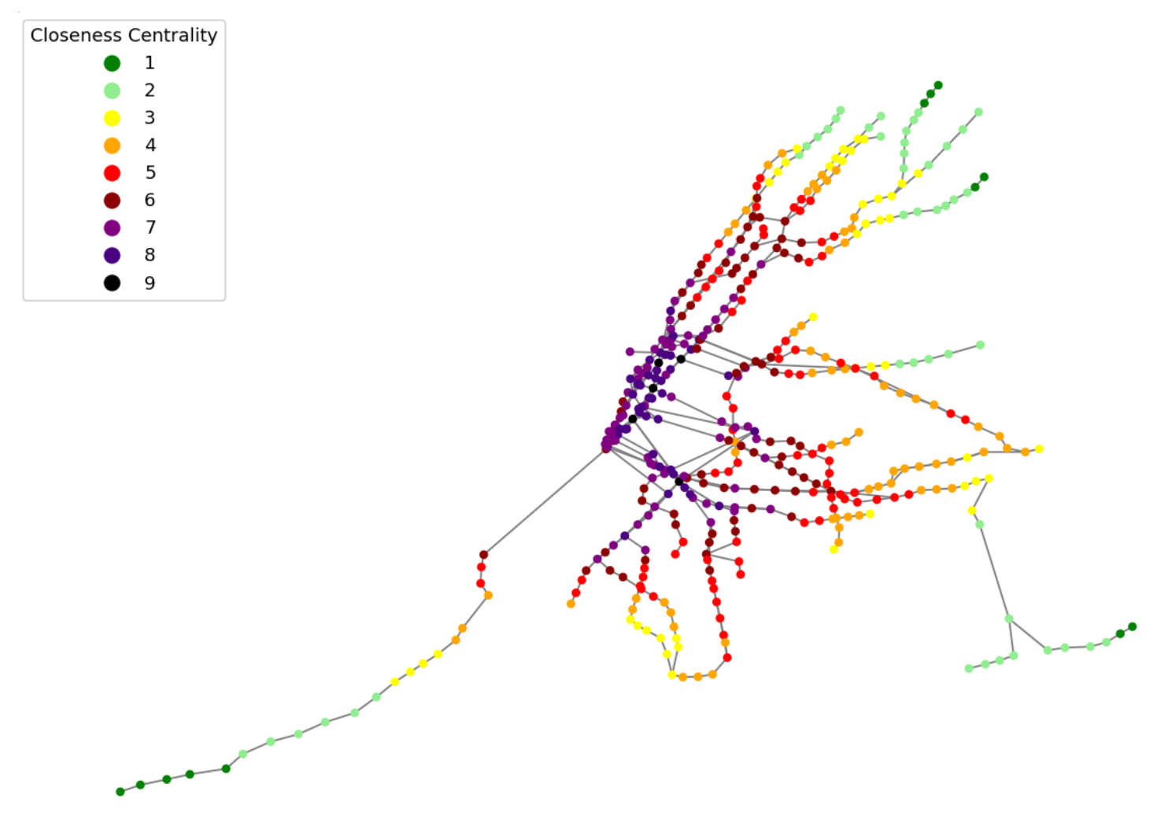
- MTA Subway Data, published by the MTA Open Data Program in 2023, included information about the name, geolocation, and subway lines running through each subway station
- NYC Public Schools Data, published by the NYC Department of Education for the 2022-23 school year, included performance metrics of each public school in Manhattan: *Student Achievement, Rigorous Instruction, Collaborative Teachers, Supportive Environment, Effective School Leadership, Strong Family-Community Ties, and Trust*

Using the Python package NetworkX, we calculated centrality measures for each subway station and generated maps to visually represent the subway network. Then, we used Google Maps' API to calculate the geolocations for all high schools and subway stations, as well as to map each high school to its nearest subway station. At the conclusion of our data processing, we had obtained each school's performance metrics and its nearest subway station's centrality measures.

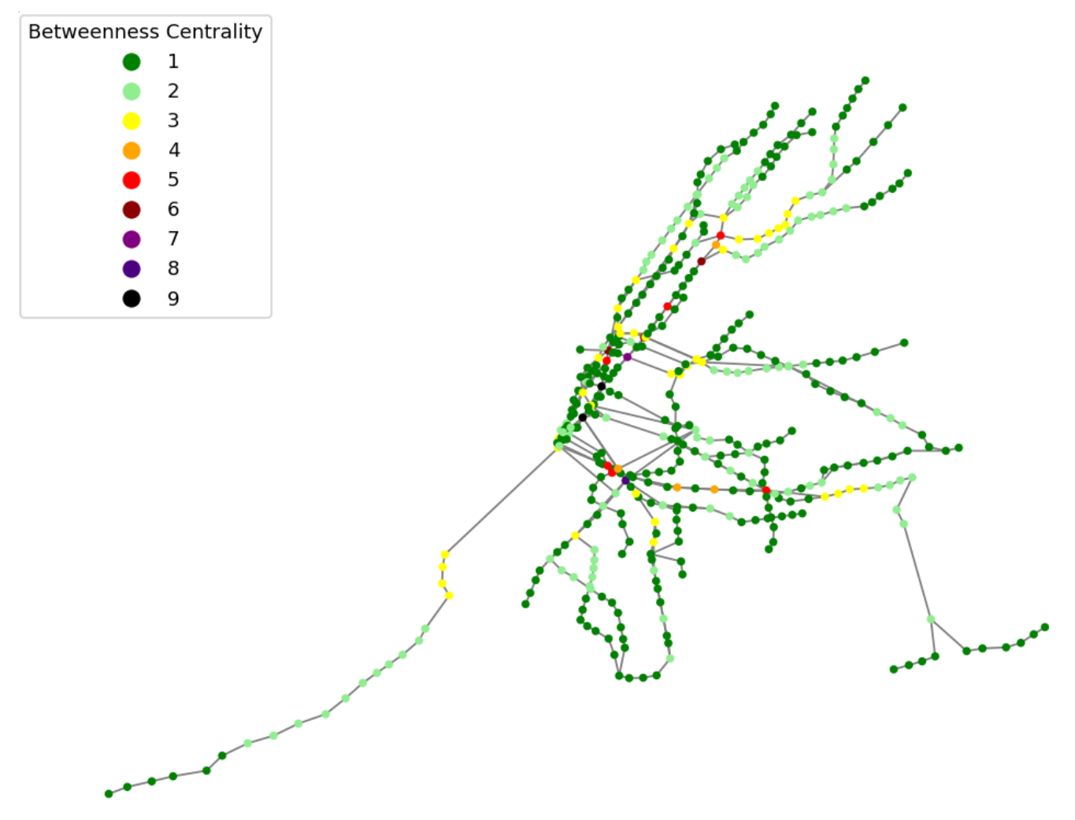
Methodology

Graph Theory

Graph theory enables complex network analysis. Node centrality is a critical concept in graph theory that indicates the importance of each node to the rest of the network. Treating each subway station as a node and the subway lines connecting subsequent stations as edges, we calculated four centrality measures for each Manhattan station based on the entire subway system: *Node Degree, Closeness Centrality, Betweenness Centrality,* and *Eigenvector Centrality.*



Closeness Centrality



Betweenness Centrality

Multiple Linear Regression

Initially, we performed Multiple Linear Regression (MLR) to analyze the relationship between subway centrality and school performance. A separate linear regression was performed for each of the 7 performance metrics.

$$\text{Performance Metric} = \beta_0 + \sum_{i=1}^{4} \beta_i \times X_i + \epsilon \qquad (1)$$

Our model satisfied several linear regression assumption checks (normality of error distribution; variance, independence, and zero mean of residuals), but the adjusted r-squared values for each performance metric were below 10%, suggesting that the MLR did not effectively capture the variation in our data. We chose to pivot from a predictive regression model to principal component analysis.

Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique that identifies correlated variables and linearly combines them into new principal components that capture information from the original components. We ran PCA on 90 rows (high schools) and 11 columns (7 performance metrics and 4 centrality measures), which resulted in the top three principal components (PCs):

- Neighbor-driven Centrality: node degree, eigenvector, and betweenness centrality, 72.6% of total variance
- Performance Metrics: all seven performance metrics, 9.12% of total variance
- Distance-driven Centrality: closeness, betweenness, and eigenvector centrality, 6.87% of total variance

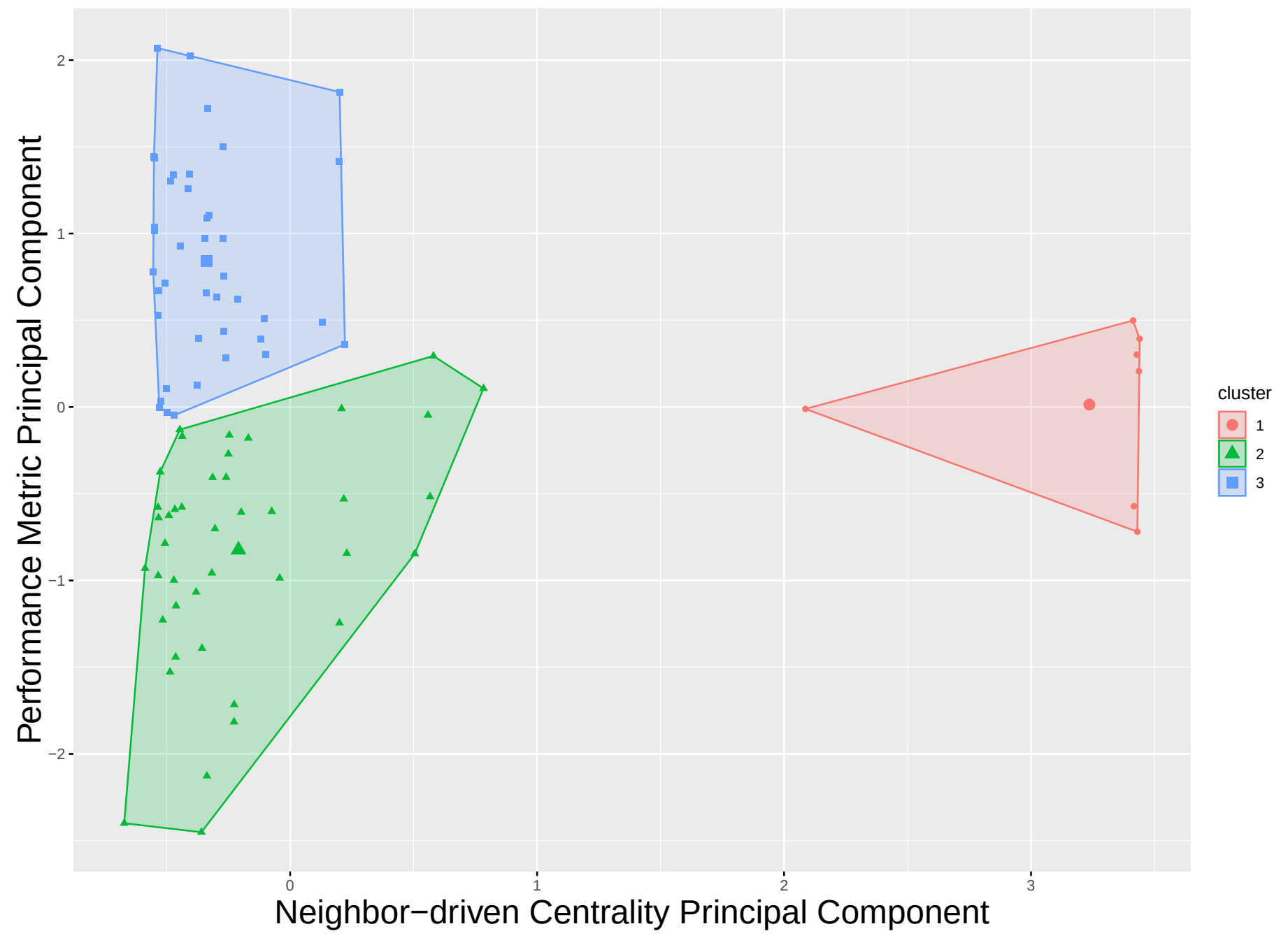
We attempted to regress each performance metric on the Centrality PCs, but the significance remained low. This result prompted us to explore clustering analysis.

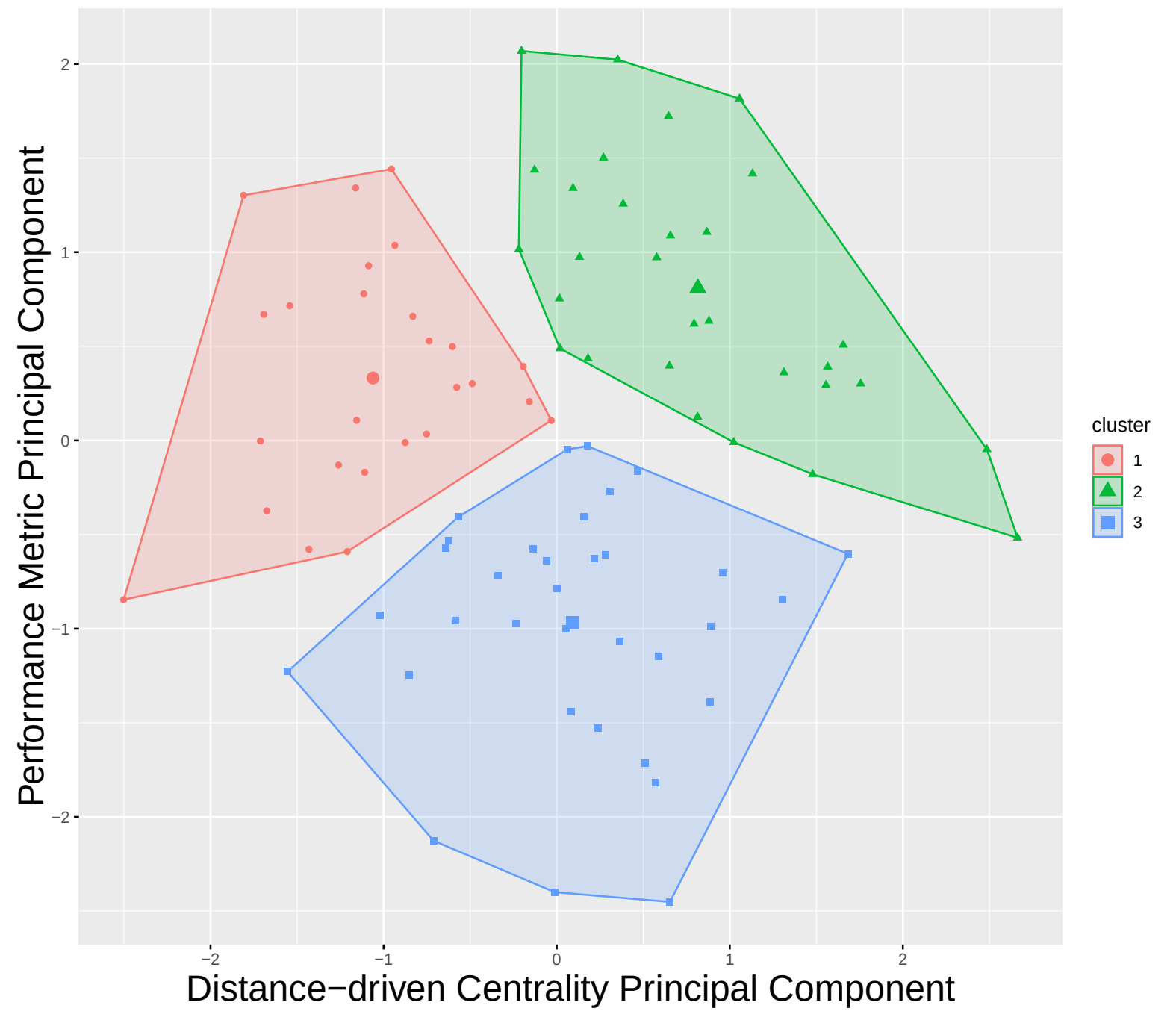
K-Means Clustering

K-means clustering is an unsupervised machine learning algorithm that partitions a dataset into *K* clusters. Using the elbow method, we formed 3 clusters based on the centrality and performance metrics PCs. After clustering, we used the original centrality and performance data to run PCA on each set of schools to identify which components were most relevant.

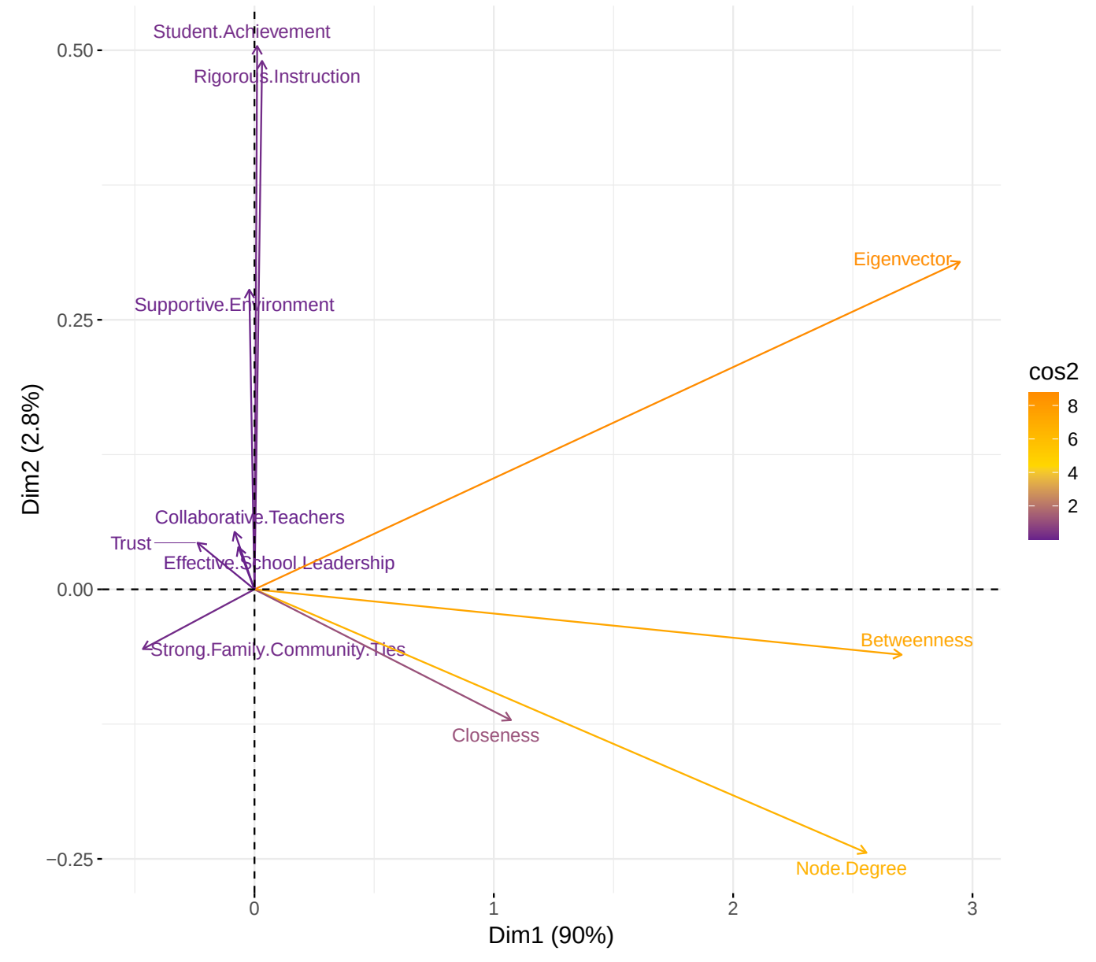
Results & Discussion

We conducted K-means clustering on both the neighbor-driven and distance-driven centrality PCs. We categorized each of the three clusters as *high* or *low centrality* in regards to the centrality PC, and *high*, *medium*, or *low performance* using the performance PC.





For both neighbor-driven and distance-driven clustering, the three clusters fell into the same centrality-performance categories: *high centrality–medium performance*, *low centrality–low performance*, and *low centrality–high performance*. To analyze these patterns further, we performed PCA on the original data for each cluster and extracted the dominant performance metrics, honoring a 55% PC loadings cutoff. Interestingly, the two centrality PCs produced the same dominant performance metrics, despite the distance-driven PC only accounting for 6.8% of the total variance, while the neighbor-driven PC captured 72.6%.



PCA: Distance-driven Cluster 1

Cluster	PC	Centrality	Performance	Dominant Performance Metric
1	Neighbor-driven	High	Med	Rigorous Instruction
2	Neighbor-driven	Low	Low	Strong-Family Community Ties
3	Neighbor-driven	Low	High	Student Achievement

Table 1. Neighbor-driven Centrality Clusters

Cluster	PC	Centrality	Performance	Dominant Performance Metric
1	Distance-driven	High	Med	Rigorous Instruction, Student Achievement
2	Distance-driven	Low	Low	Strong Family-Communities Ties
3	Distance-driven	Low	High	Student Achievement

Table 2. Distance-driven Centrality Clusters

Conclusion

Although our data didn't demonstrate the linear relationship we had initially hypothesized, using PCA and K-means clustering, we were able to reach conclusions about centrality-performance relationships for Manhattan public high schools. Namely:

- Neighbor-driven and distance-driven centrality PCs generated similar clusters, indicating that the clustering approach effectively captured centrality-performance patterns
- Within these clusters, schools associated with particularly high subway centrality did not achieve high performance, challenging our initial hypothesis
- Student Achievement is the most informative performance metric for schools with medium to high performance, whereas Strong Family-Community Ties are prominent in schools with low performance

School quality is informed by a host of factors, and school centrality and performance have a complex relationship that is worth further exploration. Expansions into other boroughs, additional centrality measures like population density around subway stations, and further exploration into non-linear models may lead to an even deeper understanding of this dynamic.

References

[1] Tri-State Transportation Campaign. How car-free is new york city?, 2017. URL `https://blog.tstc.org/2017/04/21/car-free-new-york-city/`.

[2] Sybil Derrible. Network centrality of metro systems. *PloS one*, 7(7):e40575, 2012.

[3] Dan Han and Shuping Wu. The capitalization and urbanization effect of subway stations: A network centrality perspective. *Transportation Research Part A: Policy and Practice*, 176(103815):1–29, 2023.

[4] Luis Herskovic. The effect of subway access on school choice. *Economics of Education Review*, 78(102021), 2020. doi:https://doi.org/10.1016/j.econedurev.2020.102021.

[5] Jean-William Laliberte. Long-term contextual effects in education: Schools and neighborhoods. *American Economic Journal: Economic Policy 2021*, 13(2):336–377, 2021.

Columbia Summer Undergraduate Research Experiences in Mathematical Modeling, advised by George Dragomir and Vihan Pandey