Impact of Subway Centrality on Manhattan High School Performance

Sandra Zelen, Kathy Xu, and Caroline Smyth
Columbia University in the City of New York

July 2024

Abstract

The New York City Metropolitan Transit Authority (MTA) subway system, the largest transportation agency in North America, serves more than 3 million riders daily. This study explores the impact of the MTA system on public high school performance throughout Manhattan, with the hypothesis that schools that are more central to the subway system tend to perform better academically. Prior research has applied graph theory to analyze public transit systems. Building upon these techniques, this study modeled the MTA system as a graph by treating each station as a node and the subway lines connecting adjacent stations as edges. Centrality measures were calculated for all subway stations, each public high school in Manhattan was mapped to its nearest subway station as a proxy for the school's centrality, and these centrality measures were related to the school's performance metrics through a series of predictive and explanatory modeling techniques.

For every school, multiple linear regression was used to regress each performance metric on all centrality measures. These models yielded low R-squared values, prompting a turn to principal component analysis (PCA). PCA generated three principal components (PCs): neighbor-driven centrality, distance-driven centrality, and school performance. By plotting each centrality PC against the school performance PC, three clusters of schools were identified for each graph: low centrality-high performance, low centrality-low performance, and high centrality-medium performance.

The dominant performance metrics driving each cluster revealed that low centrality-high performance schools were primarily driven by student achievement, low centrality-low performance by strong family-community ties, and high centrality-medium performance by rigorous instruction. Surprisingly, clusters from both plots shared the same dominant performance metrics, despite being derived from different centrality PCs.

Contents

1 Introduction

Only 23.4% of Manhattan households own a car, making residents of the borough highly dependent on the subway for transportation [3]. We explored how the varying strength of the MTA system throughout Manhattan shapes the borough, using centrality as a measure of that strength. Centrality is a concept in graph theory that represents the relative importance of a node to the rest of the network. Different types of centrality capture different aspects of node importance—considering how many neighbors a node has, or how far a node is from the rest of the network.

Previous research has studied how subway centrality shapes various socioeconomic outcomes including job mobility [10], racial inequities [15], 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 schools are important drivers of community success—education quality has been linked to long-term educational attainment and higher fiscal valuations of neighborhoods [11]. 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.

Our project uses a series of mathematical and statistical analysis techniques to answer the question: *does subway centrality impact the performance of public high schools in Manhattan?* We hypothesize that schools with higher centrality tend to perform better because they are more accessible via the subway.

2 Literature Review

The motivation of our study relies on the expectation that a school's performance is impacted by its surrounding urban landscape. Former literature substantiates this expectation: Jean-William Laliberté's publication in the American Economic Journal found that moving to a better neighborhood with higher quality schools makes students more likely to enroll in university after graduating high school [11]. The results of this particular study motivated us to explore school quality, as the quality of education often contributes to long-term life outcomes.

More closely related to our study, Luis Herskovic's contribution to the Economics of Education Review in 2020 studies the relationship between subway access and school choice [8]. He analyzed the 2005 inauguration of a new subway line in Santiago, Chile, and found that when travel time was reduced by public transit, parents were willing to send their children further distances for better schools. This paper strongly indicates that higher public transit connectivity impacts school choice—following this thread, we hope to identify if better transit is reflected in heightened performance indicators.

Wyczalkowski, Welch, and Pasha's 2020 publication to the Journal of Comparative Urban Law and Policy focuses on the sociological impacts of transit networks, linking the connectivity of bus and rail systems to urban race analysis in Atlanta. They combined speed, activity density, distance, and connecting power of different stops to calculate a ”connectivity index” for each transit route, then performed linear regression to estimate the impact of transit connectivity on race and poverty. They found that poorer, predominantly Black areas tended to have lower-quality public transit options [15]. Their use of linear regression to relate public transit to social outcomes proved as a potentially viable model for our work.

Diving deeper into the measure of subway centrality, Sybil Derrible evaluates the network centrality of thirty-two metro systems in the major cities across the world and highlights the importance of different centrality measures, such as betweenness and degree centrality [5]. This study inspired our project to factor multiple centrality measures into our centrality calculations, which tease out multiple angles to understanding a subway stop's ”connectedness” to the remainder of the system. Also in the realm of graph theory, a study based at Beijing Jiaotong University examined the relationship between station centrality and land allocation and housing prices. Similar to the paper from the Journal of Comparative Urban Law and Policy, this article raises nuances for the notion of centrality like train schedule and frequency data, adding an additional layer of analysis to subway centrality.

3 Data Scope and Collection

We chose to limit the geographical scope of our project to schools in Manhattan—the borough with the highest subway usage rates in New York City [3]. However, since many students commute from other boroughs to Manhattan for school, we factored the entire subway network into our centrality calculations. Furthermore, we only considered public high schools in our model, as New York City offers school buses as a transportation option for students in kindergarten through 6th grades, which likely lowers subway usage rates for elementary and middle schoolers [13]. Our analysis did not include private schools because they do not release the same performance metrics as public schools.

For data collection, we turned to two sources:

- **MTA Subway Data:** Published by the MTA Open Data Program in 2023, with information about the name, geolocation, and subway lines running through each subway station [4]. The main challenge was that certain stops situated on the same street shared the same stop name (for example, there are three locations for 86th street), thus we renamed each stop to include the street at the intersection of the subway station, so that each stop name was distinct. We removed extraneous information in the dataset such as repeated stop data for stations with multiple entrances by computing the mean latitude and longitudes of all entrances for each subway station and using the averages as the station's coordinates [7].

- **NYC Public Schools Data:** Published by the New York City Department of Education for the 2022-23 school year, which includes a dashboard with performance metrics—*Student Achievement (SA), Rigorous Instruction (RI), Collaborative Teachers (CT), Supportive Environment (SE), Effective School Leadership (ESL), Strong Family-Community Ties (SFCT), and Trust (T)* [12]. These metrics lie on a scale from 1 to 4.99 and are evaluated through surveys of families, students, and staff, and two-day observations conducted by experienced educators. We scraped the performance metrics for each Manhattan public high school using Selenium. We removed all high schools with incomplete performance information and were left with 90 high schools with 7 performance metrics each.

We then consolidated our information so that each school had its seven performance metrics and the four centrality measures of its associated station (see Table 4 and Table 5 in Appendix). We performed z-score standardization with the combined dataset so that all centrality measures and performance metrics were distributed around zero. The dataset was randomized in order to prevent biased ordering of the information. At the conclusion of our data collection process, we had obtained the geolocations of all subway stations in New York City and the performance metrics of 90 public high schools in Manhattan.

4 Methodology

Our research methodology consisted of two stages: data processing and data analysis. For the first stage, we used graph theory to model the complete subway network and determine centrality measures for each subway stop in Manhattan. Then, we mapped each high school to its nearest subway station as a proxy for the school's centrality. For the second stage, we employed a series of mathematical models relating school centrality and performance, starting with Multiple Linear Regression before moving to Principal Component Analysis and K-Means Clustering.

4.1 Graph Theory

Graph theory studies systems of nodes connected by edges. Graphs are often used to represent social networks, chemical structures, and—in our case—transit infrastructure [2]. In our model of the MTA system, we treat each subway station as a node and the subway lines connecting adjacent stations as edges. The weights of the edges represent the time in seconds between adjacent stations.

Graph Centrality Measures: There are multiple ways to calculate the centrality of a node. Our model considers four centrality measures—each of which employs a different formula—to gather a holistic understanding of a node's relative importance to the rest of the system [1].

- **Node Degree Centrality (d_i)** measures the number of edges adjacent to a given node i by summing the elements in the corresponding row (or column) in the adjacency matrix A:

  $$d_i = \sum_{j=1}^{n} A_{ij} \tag{1}$$

  Node degree is the most straightforward measure of centrality and provides a baseline of how connected a node is to the rest of the graph.

- **Eigenvector Centrality (c_j)** measures the influence of a node i within a network, calculated as a weighted sum of the centrality values c_j of neighboring nodes:

  $$c_i = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij} c_j \tag{2}$$

  Eigenvector centrality provides insight into how a subway station impacts other stations. For example, a low centrality subway stop directly connected to a high centrality stop will have higher eigenvector centrality than a low centrality subway stop only connected to other low centrality stops.

- **Betweenness Centrality (B_i)** measures the extent to which a node lies on the shortest paths between other nodes. p_{qr} is the total number of shortest paths from node q to node r, whereas $p_{qr}(i)$ is the number of those paths that pass through node i:

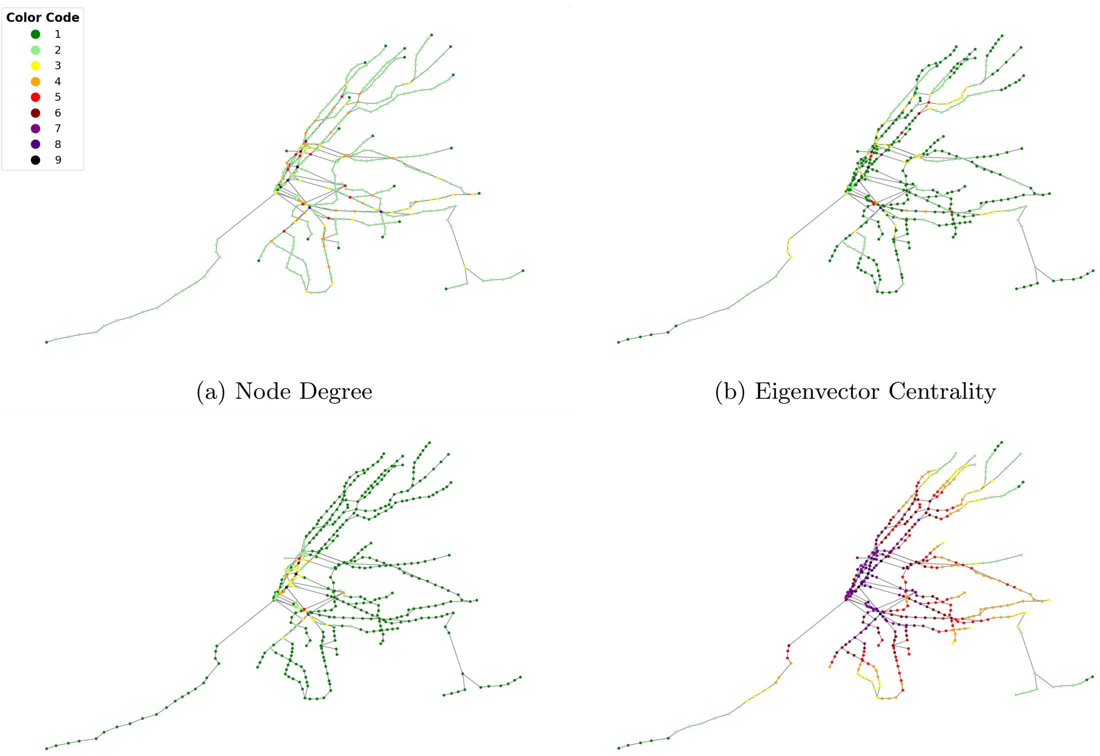
  $$B_i = \sum_{q \neq i \neq r} \frac{p_{qr}(i)}{p_{qr}} \tag{3}$$

  It provides insight into how individual nodes control the flow around a network, or which nodes' removals would impact the network the most.

- **Closeness Centrality (C_i)** measures how close a node i is to all other nodes in the graph. It is reciprocal to the sum of the shortest distances (y) between nodes i and j:

  $$C_i = \frac{1}{\sum_{j=1}^{n} y(i,j)} \tag{4}$$

  Unlike the other centrality measures, closeness centrality takes the geographical distances between subway stops into account, which influences commute time and efficiency. Stops with higher closeness centrality allow for shorter travel times to all other stops in the network.



(a) Node Degree

(b) Eigenvector Centrality

(c) Betweenness Centrality

(d) Closeness Centrality

Figure 1: Centrality Measure Graphs. Scaled from 1 (Low Centrality) to 9 (High Centrality).

Using the Python package NetworkX and geolocation data from the MTA, we modeled the subway system as a graph and calculated the four centrality measures for each subway station. We also generated color-coded graphs, which provide a visual understanding of how each centrality measure weighs each station in the subway system (see Figure 1). Finally, using Google Maps' API, a software that enables users to build and analyze custom maps, we link each school with its nearest subway station by walking distance. We consolidated our data to include each school, its performance metrics, the associated subway station, and that station's centrality measures (see Table 4 and Table 5 in Appendix).

4.2 Multiple Linear Regression

Multiple Linear Regression analysis was implemented to predict each high school performance metric based on subway centrality measures [14].We regressed each performance metric against all centrality measures, adapting Equation 5 for linear regression:

$$\begin{aligned}\text{Performance Metric} = \beta_0 &+ \beta_1 \times \text{Node Degree} \\ &+ \beta_2 \times \text{Eigenvector Centrality} \\ &+ \beta_3 \times \text{Closeness Centrality} \\ &+ \beta_4 \times \text{Betweenness Centrality} + \epsilon\end{aligned} \tag{5}$$

In order to perform multiple linear regression, we first ran a series of linear assumption tests:

- Homoscedasticity (constancy of the variance of residuals) using the Breusch-Pagan Test, resulting in $p > 0.05$ for all regressions, failing to reject constant variance of residuals.
- Autocorrelation (independence of variables) using the Durbin-Watson Test, resulting in $p > 0.05$ for all regressions, failing to reject independence of variables.
- Normality of residuals using the Shapiro-Wilk and Anderson-Darling tests, resulting in $p > 0.05$ for all performance metrics except Student Achievement, failing to reject normality of nearly all performance data.
- Calculated the means of the residuals and found means below 1.0e-16, i.e. approximately 0, for all regressions.

Our data met nearly all linear assumption tests (with the exception of normality of Student Achievement), indicating that MLR was a viable model for our data. We moved on to calculate the following statistics (see Table 4 and Table 6 in Appendix):

- R-squared and adjusted R-squared values, which helped us understand the proportion of the variance in the performance metric that is explained by the centrality measures. Both R-squared and adjusted R-squared values were below 5% for all regressions, indicating that the centrality measures were not effective in predicting performance metrics.
- Variance Inflation Factor (VIF) of each independent variable, which checks for multicollinearity between centrality measures. Closeness had a low VIF, whereas Node Degree, Eigenvector, and Betweenness Centrality exhibited problematically high VIFs (see Table 7 in Appendix).
- Error checked using Root Mean Squared Error (RMSE): $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$, where n is the number of observations, y_i is the actual value and $\hat{y}_i$ is the predicted value.
- F-Statistic, which determines whether the means between two populations are significantly different. Each F-Statistic resulted in $p > 0.05$, so we fail to reject the null hypothesis, indicating that there is no significant linear relationship between the dependent variable and the independent variables.

Although most of the linear regression assumption tests were met, the low adjusted R-squared values, failed F-statistic, and high variance inflation factors prompted us to look beyond a predictive model and consider principal component analysis.

4.3 Principal Component Analysis

Principal Component Analysis is a dimensionality reduction technique that aims to simplify data into a more concise set of features [6]. PCA identifies correlated variables and linearly combines them into principal components that each capture information from several original components. It aims to identify components that capture maximal variance amongst the points to maintain as much detail as possible. Maximized variance and minimized residuals of a principal component occur simultaneously.

Variable	**Neighbor-driven**	**Performance**	**Distance-driven**
Node Degree	0.532	0.019	-0.192
Closeness	0.162	0.044	0.61
Betweenness	0.57	-0.056	-0.529
Eigenvector	0.602	0.085	0.486
Student Achievement	0.002	0.277	0.076
Rigorous Instruction	-0.002	0.521	0.071
Collaborative Teachers	-0.019	0.43	-0.009
Supportive Environment	-0.008	0.31	-0.001
Effective School Leadership	-0.01	0.39	-0.152
Strong Family Community Ties	-0.051	0.263	-0.153
Trust	-0.022	0.369	-0.131

Table 1: Variable Loadings for Different Principal Components

We identified three principal components that capture, cumulatively, about 88% of the variance in our data (see Table 10 in Appendix). Table 1 displays the variable loadings for the three principal components. The first principal component, which represents 72.6% of variance, is loaded on primarily by node degree, betweenness, and eigenvector centrality. We chose to label this PC ”neighbor-driven centrality” because its loadings are mainly concerned with how many stops one station is connected to through subway lines. The second PC was loaded on by the seven school performance metrics and represents 9.12% of variance. The third principal component represents 6.87% of variance and, similarly to the first PC, was loaded on by centrality measures rather than performance metrics. We decided to refer to this PC as ”distance-driven centrality” because its loadings are closeness, betweenness, and eigenvector centralities.

Closeness centrality explains connectivity based on the length of subway lines and how far stops are from the rest of the network. We found that closeness centrality values features a dataset with a normal distribution (Figure 2). However, node degree, betweenness, and eigenvector centrality values follow a log-normal distribution as shown in Figure 3. The heavy right-skew in data can be explained by the subway network structure. Since most subway stations are adjacent to two other stations (the one preceding it and the one following it), there are only a few stations in the network whose centrality differs substantially.

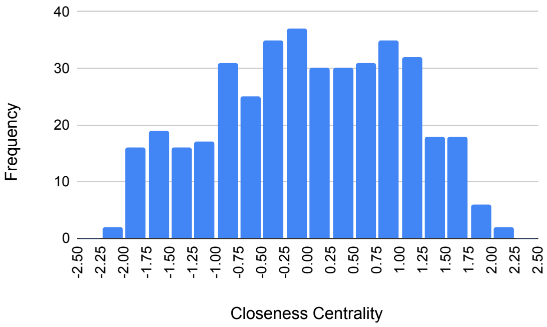


Figure 2: Closeness Centrality Distribution

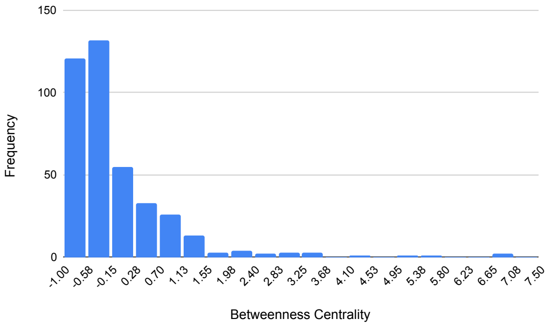


Figure 3: Betweenness Centrality Distribution

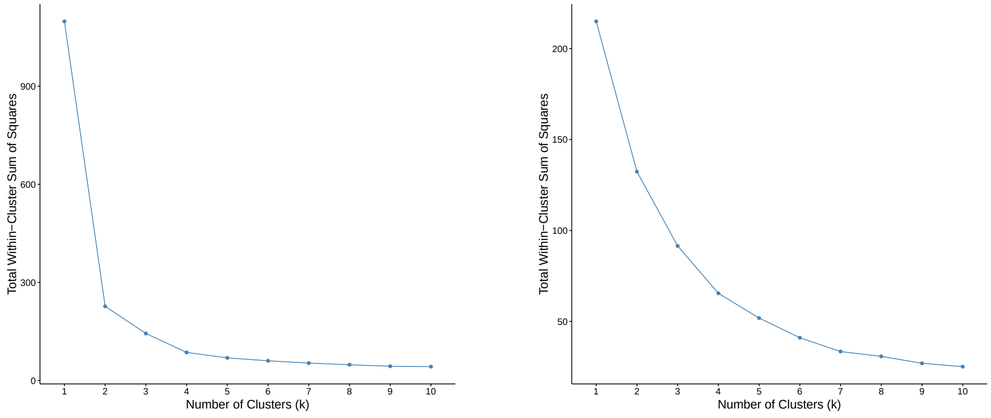
This sparked an interest in the ways that neighbor-driven and closeness-driven centrality are related to school performance. We decided to explore k-means clustering to discern potential differences in how the two centrality principal components are grouped with education quality metrics.

4.4 K-Means Clustering

K-means clustering is an unsupervised machine learning algorithm that partitions a dataset into K clusters. Using the elbow method [9], we identified that three clusters would divide the neighbor-driven graph into distinctive groups (Figure 4a). Although the distance-driven scree plot (Figure 4b) showed that six clusters would be appropriate, we decided to proceed with three both to maintain consistency and because the data was only made up of 90 columns.

We then set up two graphs with three clusters, the first with the neighbor-driven centrality principal component on the x-axis and the performance metrics principal component on the y-axis. The second graph featured the distance-driven centrality PC on the x-axis and performance metrics on the y-axis.

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 to schools with varied combinations of centrality and performance. We decided to describe our clusters' characteristics as having "high" or "low" centrality and "high", "medium", or "low" performance. The binary description of centrality was appropriate because three out of the four centrality measures follow a log-normal distribution, which doesn't provide enough scope for a "medium" category. However, the performance metrics



(a) Neighbor-driven

(b) Distance-driven

Figure 4: Elbow Method

are distributed normally and adding a ”medium” category allows for greater nuance than just two “high” and “low” options.

5 Results and Discussion

We conducted K-means clustering on both the neighbor-driven and distance-driven centrality PCs (Figure 5 and Figure 6). 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.

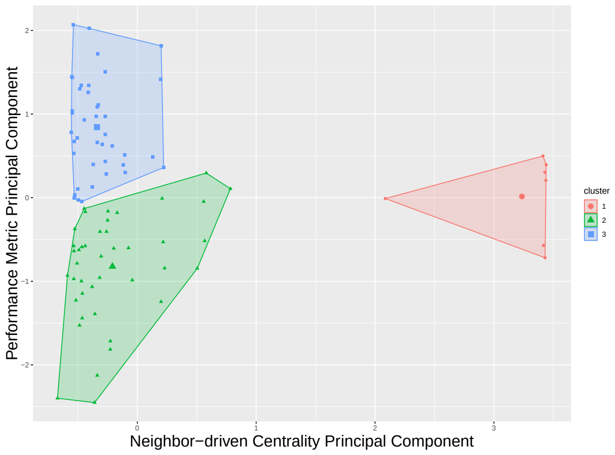


Figure 5: Neigbor-driven Clusters

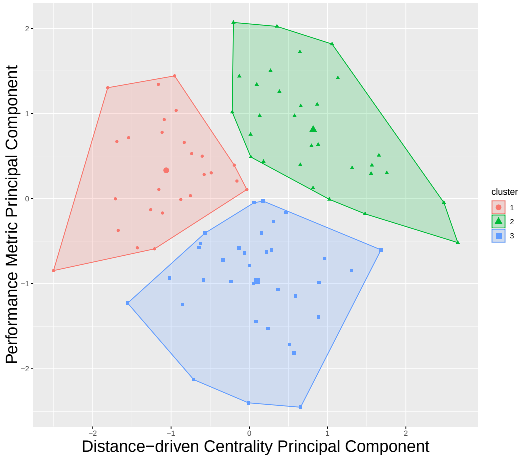


Figure 6: Distance-driven Clusters

Table 2 and Table 3 display the results of the principal component analysis on the original data of the clusters generated through k-means clustering. Our intention here was to create groups based on how central and high-performing the schools are and to deduce through PCA which variables were most important to the different groups. After labeling the clusters and matching the neighbor-driven groups to their complementary distance-driven groups, we found that each group, across centrality types, shared the same dominant performance metrics. For example, the clusters with high centrality and medium performance are both most impacted by Rigorous Instruction.

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.

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 2: Neighbor-driven Centrality Clusters

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

Table 3: Distance-driven Centrality Clusters

Interestingly, the two centrality PCs produced the same dominant performance metrics for every equivalent cluster, despite the distance-driven PC only accounting for 6.8% of the total variance, while the neighbor-driven PC captured 72.6% (see Table 10 in Appendix).The centrality-performance patterns remain consistent, regardless of whether centrality is neighbor-driven or distance-driven.

6 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

The findings inform us that quantitative performance metrics, which take into account test scores and grades, correspond to high performance. On the other hand, schools with low performance are driven by community-based performance metrics. This suggests that both quantitative and qualitative performance metrics play a significant role in determining overall school performance. Since high centrality does not drive high performance, it is worthwhile to further explore the factors which influence each cluster of high schools.

School quality is informed by a host of factors, and school centrality and performance have a complex relationship that is worth further exploration. Because New York City features such a wide array of high schools—private, public, specialized public, and charter—it could be worthwhile to examine the more limited scope of public middle schools in the city. 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] Francis Bloch, Matthew O. Jackson, and Pietro Tebaldi. Centrality measures in networks, June 2019.

[2] John Adrian Bondy, Uppaluri Siva Ramachandra Murty, et al. *Graph theory with applications*, volume 290. Macmillan London, 1976.

[3] Tri-State Transportation Campaign. How car-free is new york city?, 2017.

[4] Data.gov. Mta subway stations, 2024.

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

[6] Adam Dhalla. The math of principal component analysis (pca), 2021.

[7] Geomidpoint. Geographic midpoint calculator. http://www.geomidpoint.com/. Accessed: 2024-06-24.

[8] Luis Herskovic. The effect of subway access on school choice. *Economics of Education Review*, 78(102021), 2020.

[9] H Humaira and R Rasyidah. Determining the appropriate cluster number using elbow method for k-means algorithm. In *Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018*, Padang, Indonesia, January 2018. EAI.

[10] Sarah Kaufman, Mitchell L. Moss, Justin Tyndall, and Jorge Hernandez. Mobility, economic opportunity and new york city neighborhoods. *SSRN Electronic Journal*, 2014.

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

[12] NYC Department of City Planning. Population: Current estimates, 2024. Accessed: 2024-06-07.

[13] NYC Department of Education. Transportation guide, 2024.

[14] Mark Tranmer and Mark Elliot. Multiple linear regression. *The Cathie Marsh Centre for Census and Survey Research (CCSR)*, 5(5):1–5, 2008.

[15] Christopher K. Wyczalkowski, Timothy Welch, and Obed Pasha. Inequities of transit access: The case of atlanta, ga. *Journal of Comparative Urban Law and Policy*, 4(1):656–684, 2020.

7 Appendix

Full Term	**Abbreviation**
Student Achievement	SA
Rigorous Instruction	RI
Community Ties	CT
Supportive Environment	SE
Effective School Leadership	ESL
Strong Family-Community Ties	SFCT
Trust	T

Table 4: Performance Metric Abbreviations

School Code	**Nearest Station**	**Node Degree**	**Closeness**	**Betweenness**	**Eigenvector**	**SA**	**RI**	**CT**	**SE**	**ESL**	**SFCT**	**T**
01M539	Delancey-Essex St	-0.32	-0.355	-0.363	-0.381	4.63	3.96	3.69	4.28	4.2	4.2	4.16
01M696	Delancey-Essex St	-0.32	-0.355	-0.363	-0.381	4.58	4.46	2.82	3.96	3.77	3.77	3.92
02M047	23 St & Park Av	-0.32	1.521	-0.612	1.669	4.26	3.88	4.02	3.69	2.74	4.33	3.33
02M135	Brooklyn Bridge - City Hall	0.52	1.819	-0.205	4.099	3.31	3.72	3.06	3.28	2.82	3.21	2.98
02M139	50 St & 8th Av	0.52	0.896	-0.650	0.331	3.07	3.56	4.25	3.99	4.65	3.7	4.07

Table 5: Abbreviated List of Schools, Associated Stations, Centrality Measures, and Performance Metrics

	SA	**CT**	**RI**	**SE**	**ESL**	**SFCT**	**T**
Multiple R-Squared	0.0621	0.0438	0.0273	0.0274	0.0198	0.1237	0.0548
Adjusted R-Squared	-0.0026	-0.0221	-0.0398	-0.0397	-0.0478	0.0633	-0.0104
F-Statistic	0.9598	0.6649	0.4072	0.4085	0.2929	2.0470	0.8399
Root Mean Squared Error	1.1538	1.0706	1.1264	0.7888	1.0884	0.9467	0.7903
Variance of Residuals	0.8678	0.5889	0.7731	0.5279	0.4272	0.2353	0.1357
Independence of Variables	0.5067	0.3726	0.1158	0.2320	0.1689	0.9297	0.8560
Normality Shapiro	0.0289	0.7359	0.3678	0.6836	0.3076	0.6473	0.9137
Normality Anderson	0.0290	0.6514	0.6056	0.8927	0.4375	0.6095	0.7405
0 Mean of Residuals	6.56e-17	-7.95e-17	1.14e-17	-1.99e-17	-2.66e-17	-7.50e-18	-1.72e-16

Table 6: Multiple Linear Regression Statistical Analysis Summary

Centrality	**Variance Inflation Factor**
Node Degree	11.39
Closeness	2.54
Betweenness	10.03
Eigenvector	8.05

Table 7: Centrality Metrics VIFs

	Node Degree	**Closeness**	**Betweenness**	**Eigenvector**
Node Degree	1.00	0.52	0.92	0.88
Closeness	0.52	1.00	0.39	0.67
Betweenness	0.92	0.39	1.00	0.83
Eigenvector	0.88	0.67	0.83	1.00

Table 8: Centrality Measures Correlation Table

	SA	**RI**	**CT**	**SE**	**ESL**	**SFCT**	**T**
SA	1.00	0.60	0.20	0.59	0.13	0.07	0.17
RI	0.60	1.00	0.60	0.72	0.52	0.18	0.34
CT	0.20	0.60	1.00	0.55	0.78	0.30	0.60
SE	0.59	0.72	0.55	1.00	0.58	0.25	0.55
ESL	0.13	0.52	0.78	0.58	1.00	0.31	0.65
SFCT	0.07	0.18	0.30	0.25	0.31	1.00	0.60
T	0.17	0.34	0.60	0.55	0.65	0.60	1.00

Table 9: Performance Metrics Correlation Table

Principal Component	**Eigenvalue**	**Variance (%)**	**Cumulative Variance (%)**
Neighbor-driven	10.96	72.62	72.62
Performance	1.39	9.13	81.75
Distance-driven	1.04	6.87	88.62

Table 10: PCA Variance and Cumulative Variance