Predicting Graduation Rates for Homeless Students in New York City Public Schools

Heidi Choi, Kenny Mai, and Hope Muller

New York University

Messy Data and Machine Learning

Dr. Ravi Shroff

May 10, 2021

Equity and excellence are crucial factors for the New York City education system (Office of the Mayor, 2021; New York City Department of Education, 2021). But when we look at how the one in ten homeless students (NYSteachs, 2020) is faring, the disparities are glaring. How can our neighborhoods, administration, and public schools better serve homeless students and their families in this challenging season of their lives? Which factors in a student's career indicate if they are in the most danger of falling behind? This paper will discuss the student and school-level demographics closely correlated with high school graduation rates and recommend interventions for first and second-year high school students.

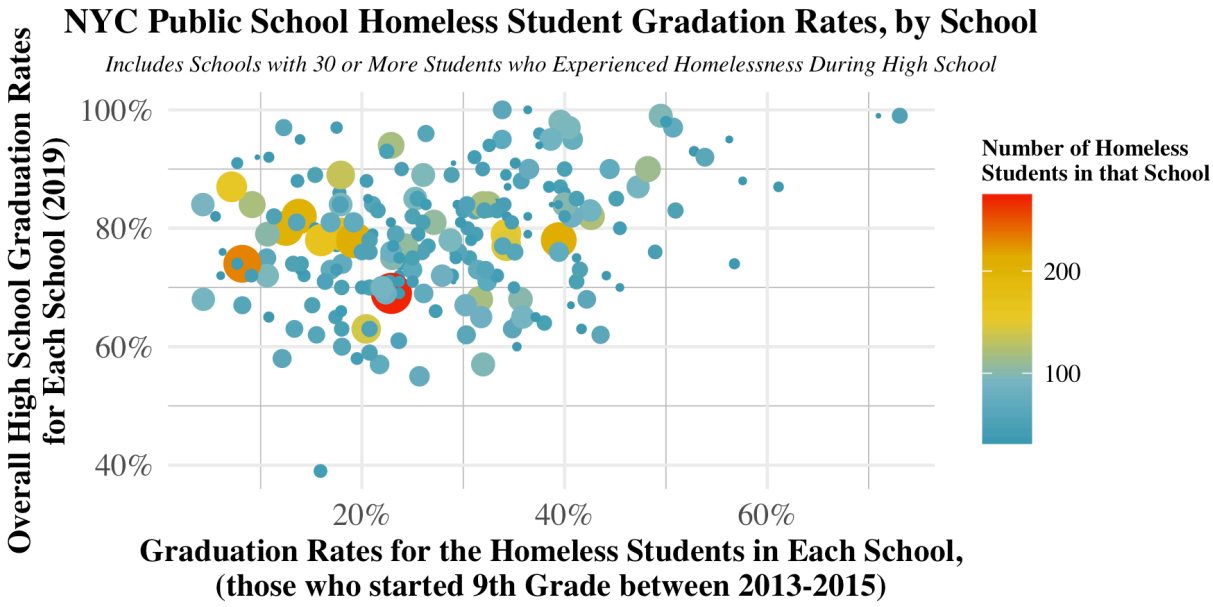
MOTIVATION

Homelessness in New York City public schools is a crisis. Kim Sweet, executive director of Advocates For Children (2019), stated that "the number of New York City students who experienced homelessness last year — 85% of whom are black or Hispanic — could fill the Barclays Center six times." Public school student homelessness is a systemic problem that must be more widely addressed in New York City.

Currently, the NYC public school graduation rate is 85% (New York State Education Department, 2016). Contrastingly, the rate for homeless students is much less, reflecting the devastating impact of homelessness on academic achievement (Civic Enterprises, 2017); the New York Times (2020) estimated it at 62%. Our project found the rates were even lower (Figure 1). And teens who don't graduate high school are 4.5

times likely to experience homelessness as adults than their peers who did graduate (Chapin Hall, 2019).

Figure 1



In addition to homelessness, there are likely many correlated factors contributing to this demographic's likelihood to graduate. The Institute for Children, Poverty, and Homelessness (2019) focused their NYC public school research on chronic absenteeism and the aspect of switching schools mid-year. Homeless students often have more significant behavior challenges at school (Hill & Mirakhur, 2019); this could also impact reaching graduation. For this project, we looked at various features for students in NYC public schools who started ninth grade between 2013 and 2015. We included a variety of individual-level and school-level variables to create prediction models for high school graduation.

DATA

The data used in this project came from The NYC Department of Education (DOE) and the National Student Clearinghouse (NSC) by way of the Research Alliance for NYC Schools housed in NYU Steinhardt. While the school-level data from the DOE are public, the Research Alliance de-identifies, aggregates, and stores the student-level data, adhering to all FERPA regulations.

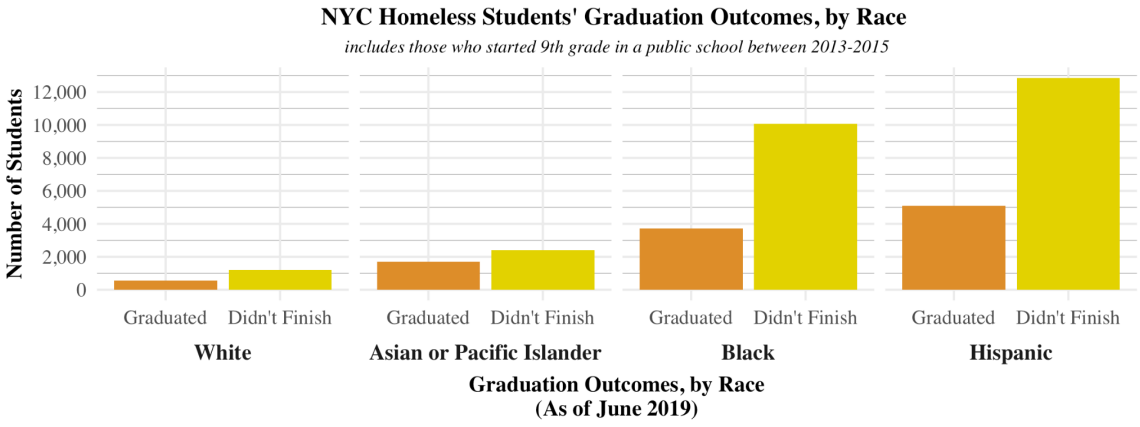
The Research Alliance formats initial student-level data annually. Every academic year, a single student's data are collected if the student enrolled in the public school system. In addition to basic demographic information, these student-level data include their enrollment status, the number of days missed due to suspensions or absences, their race/ethnicity, whether or not they've spent time in a shelter, whether they are an English language learner, whether they received special education services from the city, and the school(s) they attended that year. We combined and aggregated the student-level data from each year, and the final features used in the dataset were binary, character, continuous, or percentage-based values.

The DOE compiled many essential school-level variables in the 2019 School Quality Reports. They calculated the total school enrollment, average graduation rates, average test scores, longevity of the principal, teacher retention, racial percentages within the school, student and teacher absenteeism, school percentage of college attendees, and characteristics of the school voted on by parents. The final dataset is a compilation of both datasets.

The NYC DOE's Special Education Students Information System keeps track of students in NYC who received any special services from the DOE -- including information about those who lived in temporary housing in any specific year. As such, this study only includes students who ever experienced homelessness during their NYC school career between 2013 and 2019.

To create a consistent set of students for the predictive model, we further subsetted the student data. We included all students who began their first year of ninth grade in the school system between 2013 and 2015 and dropped out or graduated by 2019 unless the administration noted they moved out of New York City. The resulting dataset contained 19,975 students; 29.5% graduated within four to six years. Figure 2 shows graduation outcomes (graduated or didn't finish) by race for students in our resulting dataset.

Figure 2



Tables 1 and 2 show some demographic information about the students in our final dataset, which we thought were contextually important to this study. The majority of

students self-identified as Black or Hispanic (Table 1), and the majority's language was either English or Spanish (Table 2). Asian students who experienced homelessness were less likely to appear in a shelter, repeat a grade, or receive special education services and more likely to graduate from high school than their homeless peers of other racial groups.

Table 1: Features of Interest, Aggregated by Race

Race / Ethnicity	Number of Homeless Students	Graduated	Repeated Grade 9 or 10	Lived in a Shelter	Received Special Education
American Indian/Alaskan Native	117	29.1 %	23.9 %	40.2 %	26.5 %
Asian/Pacific Islander	2,253	39 %	9.3 %	4.3 %	4.7 %
Hispanic	9,601	26.1 %	18.7 %	31.7 %	21.2 %
Black, Non-Hispanic	7,271	24.6 %	21.3 %	52.8 %	26.5 %
White, Non-Hispanic	975	28.4 %	15.5 %	21.9 %	19.4 %
Unknown/Unspecified	120	19.5 %	15.3 %	41.5 %	17.8 %
Mixed Race	33	31.3 %	21.9 %	46.9 %	21.9 %

Table 2: Top 10 Languages Spoken at Home

Home Languages	Number of Homeless Students	Percentage
English	10,822	54 %
Spanish	6,226	31 %
Bengali	532	2.7 %
Mandarin	519	2.6 %
Cantonese	255	1.3 %
Any Chinese	246	1.2 %
Haitian Creole	230	1.2 %
Arabic	204	1 %
French	166	.8 %
Russian	131	.7 %

METHODS

The first of our two machine learning methodologies was the random forest, which is an extension of decision trees with bootstrap aggregation. Our research question aims to predict high school graduation for students who experienced homelessness during their time in the NYC public school system, with the intent of informing educational policy reforms and intervention programs. To that end, we wanted to select a method that would limit overfitting the data and reduce errors due to bias and variance.

Since our outcome was binary ("graduate" = 1/0), and our data had many different features, we used a random forest to see which features best predict whether an individual graduates from high school without worrying too much about bias. Moreover, because random forests use a random subset of features in many different decision trees, we thought this was an appropriate method to predict high school graduation for students who have experienced homelessness during their time in the NYC public school system.

We also used cross-validated LASSO models for student-level data and a multi-level LASSO model that incorporated school-level data. Our data contained several student- and school-level predictors that may be important features to consider. For example, is a student who has experienced homelessness more likely to graduate from high school if they attended a high school with a higher overall graduation rate? What if they went to a high school with a high teacher retention rate? We thought that the LASSO models would perform well, especially because they are more interpretable due to producing sparser models. Additionally, since LASSO prevents overfitting and performs

some feature selection by reducing some coefficients to zero, we thought it would select the most necessary predictors and created models from these predictors.

RESULTS

Random Forest

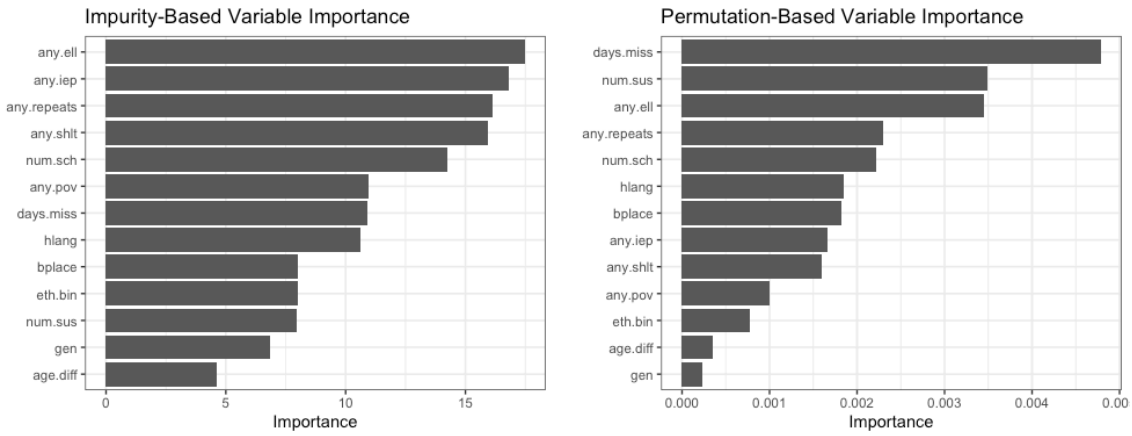
We fitted a random forest model to predict high school graduation using the *ranger* package in R. We tuned the model (via the *caret* package) to find the best hyperparameters (i.e., number of trees, number of features, minimum node size). Table 3 shows the variables included in the model and their order of importance, which was determined by the average total reduction of the impurity for a given variable across all trees, otherwise known as "impurity". This is contrasted with the "permutation"-based approach, where each out-of-bag sample is passed down each tree, and the resulting prediction accuracy is recorded. The values for each variable are then randomly permuted, alongside the prediction accuracy. The resulting decreases in accuracy are averaged over all the different trees for each feature, where the features with the largest average decrease in accuracy are interpreted as more important.

Table 3: Selected Features for Random Forest (Impurity vs. Permutation)

Grade 9 and 10 Variables Included in Random Forest	Order of Variable Importance Based on Impurity	Order of Variable Importance Based on Permutation
Received English Language Services (ELL) (*any.ell*)	1	3
Received Special Education Services (IEP) (*any.iep*)	2	8
Repeated a Grade (*any.repeats*)	3	4
Ever Resided in Shelter (*any.shlt*)	4	9
Number of Schools Attended > 1 (*num.sch*)	5	5
Family Ever Lived Under Poverty Line (*any.pov*)	6	10
School Days Missed >= 10 (*days.miss*)	7	1
Home Language not English (*hlang*)	8	6
Birthplace not Domestic US (*bplace*)	9	7
Ethnicity was Black or Hispanic (*eth.bin*)	10	11
Number of Suspensions > 0 (*num.sus*)	11	2
Gender (*gen*)	12	13
Age Difference from Expected Grade-Age > 0 (*age.diff*)	13	12

Based on this comparison, we can see that repeating a grade, ever residing in a shelter, and number of schools all appear in the top 5. Figure 3 shows the breakdown of these two comparisons by values. The impurity-based model predicted out-of-sample high school graduation probabilities between .08 and .43, and the AUC for the model was 0.715. The permutation-based model predicted out-of-sample high school graduation probabilities between .07 and .45, with the AUC at 0.726.

Figure 3: Variable Importance Comparison



LASSO

The cross-validated LASSO model weaned out several variables when creating a student-level model. It deemed some features less necessary in predicting high school graduation outcomes. Then, we ran the multi-level LASSO feature selection with individual and school-level variables. Through the AUC, we chose an ideal tuning parameter. The table below shows the features selected and their importance. One of the most critical features in the model was the percent of school days the student missed. In our data, eighty-six percent of the homeless students attended school every day, however, 1 out of every 40 students participated in a month or less of school. As seen in Figure 4, when students miss the entire school year due to suspension or other absences, it significantly impacts their graduation outcomes. The models found that an individual's gender, race, and birthplace were not important features in predicting graduation with this data. One of the most important features in the model with school-level variables were the teacher attendance rates with a positive value and the percentage of special education

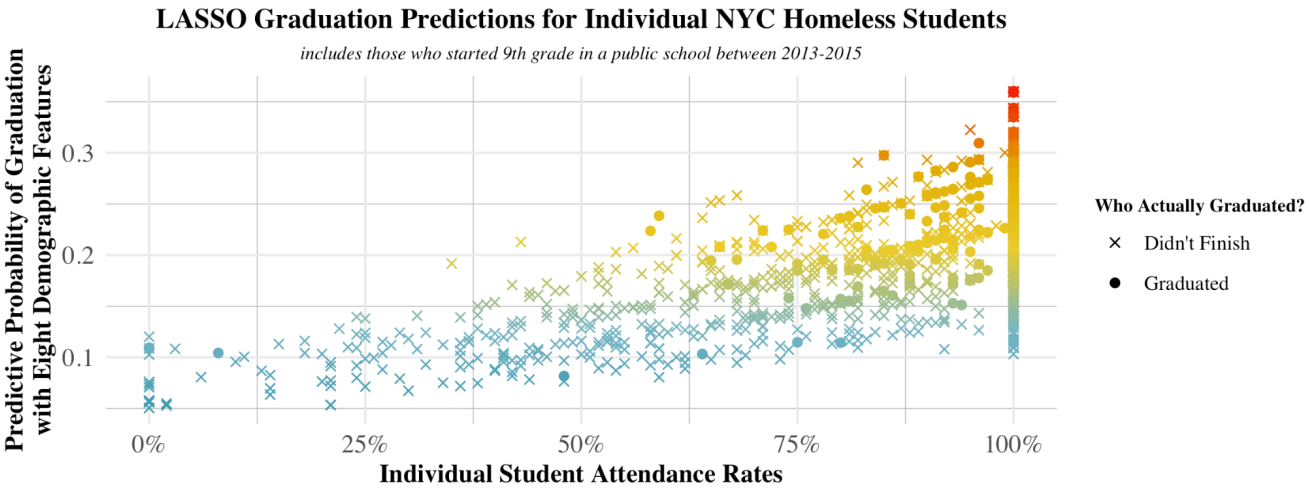
students contained separately from the mainstream classroom, with a negative coefficient. Schools in our data have up to 15% of their students in a self-contained program. The third crucial school-level feature was school-wide graduation rate, with a positive coefficient. Schools that have higher graduation rates correlated with higher graduation rates for homeless individuals at that school.

Table 4: Selected Features for LASSO

Grade 9 and 10 Variables Entered Into LASSO	Student-Level Variables Selected by CV LASSO	Student and School-Level Variables Selected by LASSO
Percent of School Days Missed	1	3
Number of Schools Attended	2	
Home Language (English or Other)	3	
Ever Resided in Shelter	4	9
Received Special Education Services	5	10
Received English Language Services	6	14
Age Difference from Expected Grade-Age	7	8
Repeated a Grade	8	7
Family Self-Reported to be Under Poverty Line		4
Moved Mid-Year		12
Mean Number of Suspensions Per Year		13
Teacher Attendance Rate (School-level)		1
Percent of Students in Self-Contained Classes		2
Schoolwide Graduation Rate		5
Percent of Students that Attended at least 3 Semesters of College (School-level)		6
Perception of Rigorous Instruction (School-level)		11
Percent of School with Special Education Services		15

The LASSO model with only student-level features predicted out of sample high school graduation probabilities between .06 and .33. The LASSO model that also included school-level features had a range of .01 to .44 probability of graduating. These aligned with the high school graduation rates for our dataset. The student-level model had an AUC of .718. The model with student and school-level features had an AUC of .734.

Figure 4



LIMITATIONS

There were several limitations in our data. In addition to missing data from the source at the DOE, we subsetted the student data due to inconsistent data for valid predictive models. The NYC DOE does not collect data on students before or after they move into the NYC public education system. This missing data cannot be imputed reliably, so we did not include students who did not start their school careers in NYC. This presented some additional limitations in what kinds of models we could fit, and how far we could interpret them.

The models intended to predict whether homeless students in grade 9 or 10 were at high risk of not graduating. We included students who began 9th grade for the first time between 2013-2015. The summary variables included their grade 9 and 10 data.

One of the primary limitations of the random forest model was computational time and resources. Training and tuning the model with different numbers of trees, features, data points, splitting rules, and nodes was complex and time consuming. (While tuning the model for better hyperparameters wasn't strictly necessary, we thought it would provide additional robustness to our modeling procedure.) The results are far less interpretable with random forest than logistic regression, and extracting and understanding the "important" variables for interpretation was far more cumbersome than initially expected. With machine learning models like random forest, we are oftentimes more concerned with prediction accuracy than interpreting the relationships between the variables so the coefficient interpretability is not the priority.

Additionally, there are a few limitations to using LASSO for predictive modeling. If a dataset has highly correlated variables, LASSO will usually remove one of them. Another notable limitation was that the coefficients do not tell a story in the way they do with a multi-level general logistic regression model. It is not possible to compare the coefficient value to the change of the outcome variable.

IMPLICATIONS AND CONCLUSION

With our prediction models, schools and communities could gain a bigger vision for homeless sophomores who may be unlikely to graduate. This moment would be one

where community services, social workers, teachers, and families could step in and help students map out their goals and set-up wrap-around support systems. These supports could ensure personal supports, reliable internet access, safe places to do schoolwork, and a place for sleep and hygiene.

Moreover, it is worth noting again that students who have experienced homelessness or move around in the shelter system are often at a disadvantage in school. They have to re-adjust every time they move, which interferes with social development and can result in students "acting out" in school (ICPH, 2013). Using these models as a foundation, we envision administrators and teachers preemptively reaching out to these students to provide the support necessary for these students to successfully integrate into their schools. This kind of intervention could be folded into existing after-school programs, where new or incoming students are "flagged" when they are coming in from a different school during the school year. A dedicated teacher or school counselor could then take on the role of welcoming the student to the school, so that the student gets a sense of immediate belonging.

The LASSO model demonstrated that school attendance is an important feature in predicting high school graduation. Although truancy may not receive the attention it needs in our school systems, most communities do try to address this with their limited budgets. Washington DC public schools spent $3.5 million to create a truancy intervention program across the district, alongside seven community agencies (Chandler, 2014). When a student missed more than five days of school without an excuse, a school

support team met with the family to create an attendance plan. Instead of looking at absences as something to punish, they addressed it as an issue where the family needs help. During this program, they saw their chronic absenteeism decrease from 27% to 18% in one year.

Some communities offer effective housing programs for their homeless youth—the Roadmap to Graduation Program in Michigan pairs homeless seniors in high school with a mentor family, and as of 2021 they report a 100% graduation and employment rate. The Homeless Youth Initiative in Virginia offers safe and stable housing for high school students who don't receive parental support (Civic Enterprises, 2017). Schools like Poway High School in Los Angeles installed showers and laundry facilities for 14% of their students, who do not have a stable living environment (Brennan, 2019). This affords students privacy and dignity during their time in school.

Both the random forest and LASSO models showed whether the student repeated grade 9 or 10 as an important feature in predicting high school graduation. After grade four, we know that repeating a grade usually negatively impacts students (Hughes, 2014). There is limited evidence-based research on successful grade retention alternatives, but many ideas exist. Chicago public schools implemented a practice that extends learning time for students without retaining them in their prior grade. This resulted in significant gains associated with their summer school intensives for third, sixth, and eighth grade students (Roderick, 2003).

Other kinds of integrated intervention programs could partner with existing mentorship programs to provide students with mentors and leaders who can provide critical guidance for especially vulnerable students. One such program already exists in New York City, called iMentor, which matches high school students with college-educated mentors that provide guidance and support to students throughout their high school and college years (iMentor, 2021).

We also note another somewhat ancillary impact of our models - they should be used to determine required resources for teachers and administrators. These resources would would allow school staff the bandwidth and support to take on the additional responsibilities of supporting these students. Public school teachers have long been under-resourced - especially in under-served neighborhoods (Smedley, Stith, Colburn et al., 2001). We hope that these models may serve to further quantify the severity and understated pervasiveness of the disparity in high school graduation rates for vulnerable students, such that policymakers can provide these schools with the resources they need.

Future studies could look at college enrollment, and factors which are predictive of college enrollment for this vulnerable population of students, and then structure targeted intervention programs around the results. Ultimately, we need to move toward addressing this disparity in new, more innovative ways that prioritize the health, safety, and wellness of our high school students, and we hope that our models serve to begin that important conversation.

References

Advocates for Children of New York. (2019). New Data Show Number of NYC Students who are Homeless Topped 100,000 for Fourth Consecutive Year. Retrieved from: https://www.advocatesforchildren.org/node/1403

Brennan, D. (2019, Dec. 26). *Poway school will add shower and laundry facilities for homeless students.* Los Angeles Times. Retrieved from: https://www.latimes.com/california/story/2019-12-26/abraxas-high-school-in-poway-adding-shower-laundry-for-homeless-students

Chandler, M. (2014, Sept. 9). D.C. anti-truancy program is getting more kids to school. Washington Post. Retrieved from: https://www.washingtonpost.com/local/education/dc-anti-truancy-program-is-getting-more-kids-to-school/2014/09/09/e83c7e12-3778-11e4-9c9f-ebb47272e40e_story.html

Chapin Hall at the University of Chicago. (2019). New Study Shows Youth Who Experience Homelessness Less Likely to Attend College. Retrieved from: https://www.chapinhall.org/news/new-study-shows-youth-who-experience-homelessness-less-likely-to-attend-college/

Civic Enterprises, Ingram, E. S., Bridgeland, J. M., Reed, B., Atwell, M.. (2017). Hidden in Plain Sight: Homeless Students in America's Public Schools. Retrieved from: https://eric.ed.gov/?id=ED572753

Hill, K. & Mirakhur, Z. (2019). Homelessness in NYC Elementary Schools: Student Experiences and Educator Perspectives. Research Alliance for New York City

Schools. Retrieved from: https://steinhardt.nyu.edu/research-alliance/research/publications/homelessness-nyc-elementary-schools

Hughes, J. N., Cao, Q., West, S. G., Allee Smith, P., & Cerda, C. (2017). Effect of retention in elementary grades on dropping out of school early. *Journal of school psychology*, *65*, 11–27. Retrieved from: https://doi.org/10.1016/j.jsp.2017.06.003

iMentor. (2021). https://imentor.org/

Institute for Children, Poverty, and Homelessness. (2013). An unstable foundation: Factors that impact educational attainment among homeless children. NYC: ICPH, USA.

Institute for Children, Poverty, and Homelessness. (2019). School Instability Factors. Retrieved from: https://www.icphusa.org/reports/school-instability-factors/#overview

New York City Department of Education. (2021). https://www.schools.nyc.gov/about-us/vision-and-mission

New York State Education Department. (2016). State Education Department Releases 2016 Cohort High School Graduation Rates. Retrieved from: http://www.nysed.gov/news/2021/state-education-department-releases-2016-cohort-high-school-graduation-

rates#:~:text=The%20State%20Education%20Department%20today,percent%20for%20the%202015%20cohort.

NYSteachs. (2020) Data on Student Homelessness in NYS. Retrieved from: https://nysteachs.org/topic-resource/data-on-student-homelessness-nys/

Office of the Mayor. (2021). The Official Website of New York City. https://www1.nyc.gov/office-of-the-mayor/index.page

Roderick, M., Engel, M., Nagaoka, J. (2003). Ending Social Promotion in Chicago: Results from Summer Bridge. Chicago: Consortium on Chicago School Research. Retrieved from: https://consortium.uchicago.edu/publications/ending-social-promotion-results-summer-bridge

Shapiro, S. (2020, Sept. 9). *The Children in the Shadows: New York City's Homeless Students*. New York Times. https://www.nytimes.com/interactive/2020/09/09/magazine/homeless-students.html

Smedley B.D., Stith A.Y., Colburn L., et al. Institute of Medicine. (2001). Inequality in Teaching and Schooling: How Opportunity Is Rationed to Students of Color in America. The Right Thing to Do, The Smart Thing to Do: Enhancing Diversity in the Health Professions: Summary of the Symposium on Diversity in Health Professions in Honor of Herbert W. Nickens, M.D..Washington (DC): National Academies Press. Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK223640/