PASSNYC

Tiffany Zhu

Contents

1 Overview

PASSNYC is a not-for-profit organization that facilitates a collective impact that is dedicated to broadening educational opportunities for New York City's talented and underserved students. New York City is home to some of the most impressive educational institutions in the world, yet in recent years, the City's specialized high schools - institutions with historically transformative impact on student outcomes - have seen a shift toward more homogeneous student body demographics.

PASSNYC uses public data to identify students within New York City's under-performing school districts and, through consulting and collaboration with partners, aims to increase the diversity of students taking the Specialized High School Admissions Test (SHSAT). By focusing efforts in under-performing areas that are historically underrepresented in SHSAT registration, we will help pave the path to specialized high schools for a more diverse group of students.

2 Problem Statement

PASSNYC and its partners provide outreach services that improve the chances of students taking the SHSAT and receiving placements in these specialized high schools. The current process of identifying schools is effective, but PASSNYC could have an even greater impact with a more informed, granular approach to quantifying the potential for outreach at a given school. Proxies that have been good indicators of these types of schools include data on English Language Learners, Students with Disabilities, Students on Free/Reduced Lunch, and Students with Temporary Housing.

Part of this challenge is to assess the needs of students by using publicly available data to quantify the challenges they face in taking the SHSAT. The best solutions will enable PASSNYC to identify the schools where minority and underserved students stand to gain the most from services like after school programs, test preparation, mentoring, or resources for parents.

Submissions for the Main Prize Track will be judged based on the following general criteria:

- Performance - How well does the solution match schools and the needs of students to PASSNYC services? PASSNYC will not be able to live test every submission, so a strong entry will clearly articulate why it is effective at tackling the problem.
- Influential - The PASSNYC team wants to put the winning submissions to work quickly. Therefore a good entry will be easy to understand and will enable PASSNYC to convince stakeholders where services are needed the most.
- Shareable - PASSNYC works with over 60 partner organizations to offer services such as test preparation, tutoring, mentoring, extracurricular programs, educational consultants, community and student groups, trade associations, and more. Winning submissions will be able to provide convincing insights to a wide subset of these organizations.

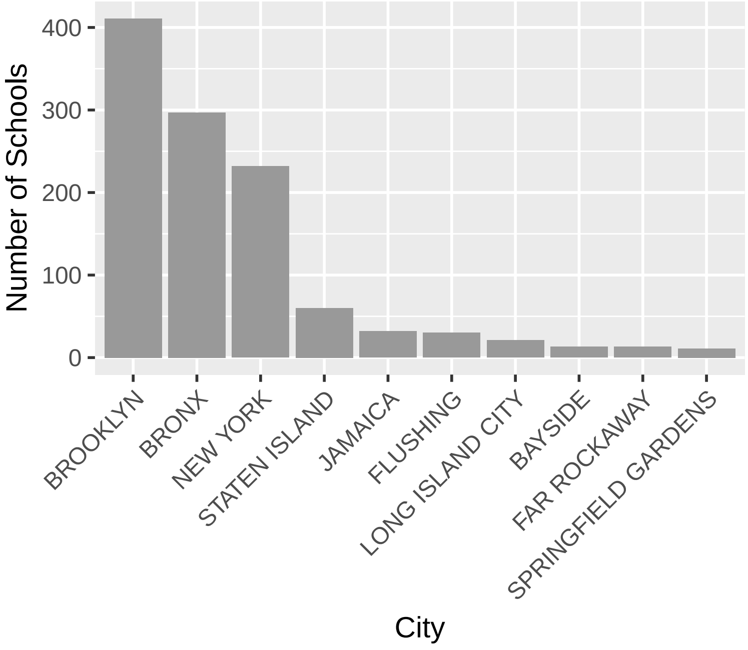
3 Loading Data

```
setwd("~/Desktop/other/data-science-for-good")
school_data <- read.csv('2016 School Explorer.csv')
shsat_data <- read.csv('D5 SHSAT Registrations and Testers.csv')
```

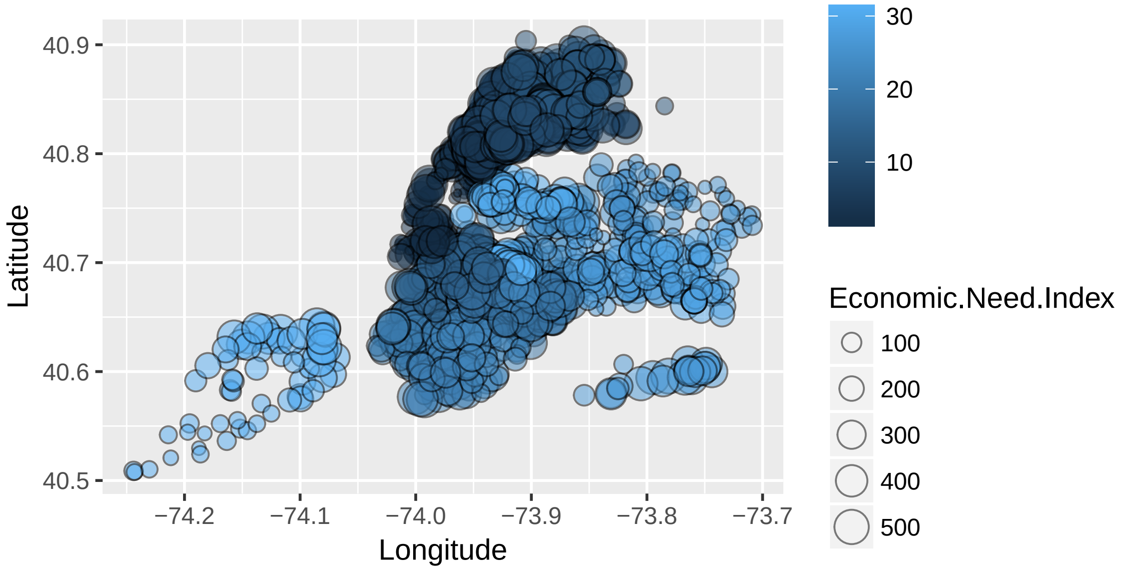
There are 1272 observations and 161 variables in `school_data`. There are 140 observations and 7 variables in `shsat_data`.

4 Data Exploration

Looking at top cities with the most schools represented in this data set:

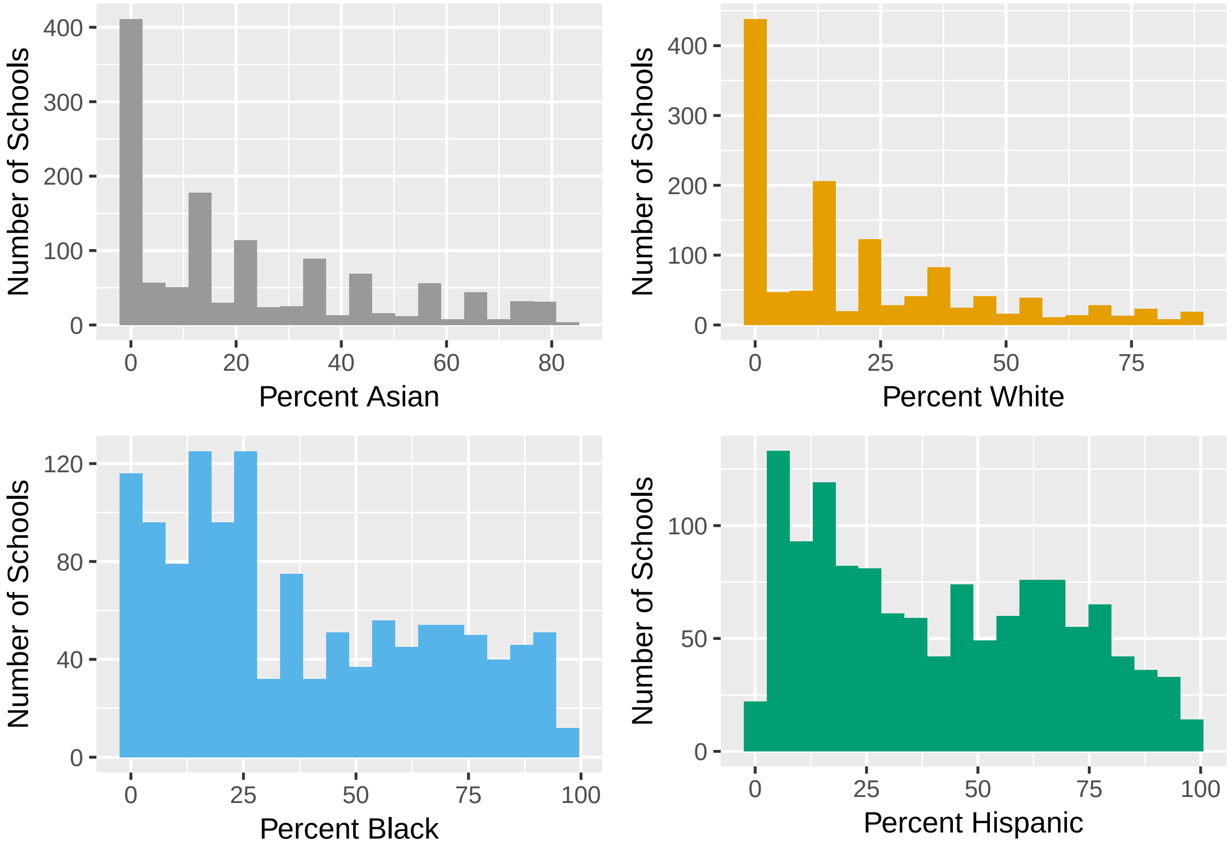


4.1 Location and Economic Need

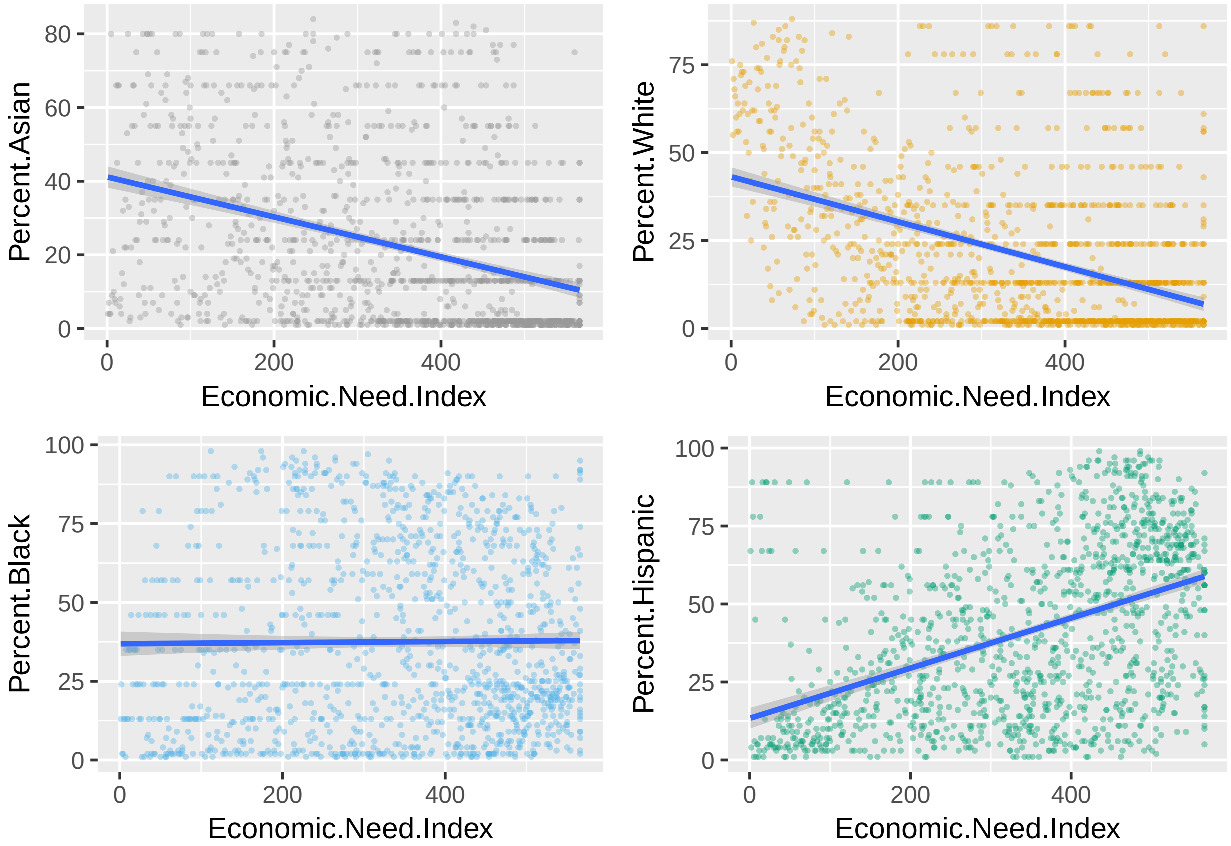


4.2 Race

4.2.1 Race and Number of School



4.2.2 Race and Economic Need



4.3 ESL Students

4.4 Student Attendance Rate

4.5 School Income Estimate and Economic Need Index

4.6 Support from the Community and School

Variables to look at:

- Community School,
- Collaborative Teachers,
- Supportive Environment,
- Effective School Leader Ship,
- Strong Family,
- Trust

5 Models

5.1 Setting Up

5.1.1 What are we going to predict?

We want to be able to predict the number of students who will (1) register for the SHSAT and (2) actually take the SHSAT. To train our model, we can use the data from `shsat_data` (more specifically, the enrollment number, number of student registered/took the SHSAT) to know the percentage of students who do both (1) and (2). Our dependent variable (the variable we are trying to predict) will be this percentage since in our testing data, we do not know the enrollment numbers. We will call these variables `percentage_registered` and `percentage_taken`. Thus,

`percentage_registered`

$$= \frac{\texttt{Number.of.students.who.registered.for.the.SHSAT}}{\texttt{Enrollment.on.10.31}}$$

and,

`percentage_taken`

$$= \frac{\texttt{Number.of.students.who.took.the.SHSAT}}{\texttt{Enrollment.on.10.31}}$$

5.1.2 How do we use the data given?

There are 21 schools in both `school_data` (characteristics about the schools) and `shsat_data` (data about schools whose students enrolled in SHST). We can use these schools to create a model.

These schools are:

	Location.Code	School.Name
1	05M046	P.S. 046 ARTHUR TAPPAN
2	05M123	P.S. 123 MAHALIA JACKSON
3	05M129	P.S. 129 JOHN H. FINLEY
4	05M148	EAGLE ACADEMY FOR YOUNG MEN OF HARLEM
5	05M161	P.S. 161 PEDRO ALBIZU CAMPOS
6	05M286	URBAN ASSEMBLY ACADEMY FOR FUTURE LEADERS
7	05M302	KAPPA IV
8	05M362	COLUMBIA SECONDARY SCHOOL
9	05M499	FREDERICK DOUGLASS ACADEMY
10	05M514	NEW DESIGN MIDDLE SCHOOL
11	05M670	THURGOOD MARSHALL ACADEMY FOR LEARNING AND SOCIAL CHANGE
12	84M065	DEMOCRACY PREP ENDURANCE CHARTER SCHOOL
13	84M284	HARLEM CHILDREN'S ZONE PROMISE ACADEMY 1 CHARTER SCHOOL
14	84M336	KIPP INFINITY CHARTER SCHOOL
15	84M341	HARLEM CHILDREN'S ZONE PROMISE ACADEMY II CHARTER SCHOOL
16	84M350	DEMOCRACY PREP CHARTER SCHOOL
17	84M384	SUCCESS ACADEMY CHARTER SCHOOL - HARLEM 2
18	84M388	ST. HOPE LEADERSHIP ACADEMY CHARTER SCHOOL
19	84M481	DEMOCRACY PREP HARLEM CHARTER SCHOOL
20	84M709	HARLEM VILLAGE ACADEMY CHARTER SCHOOL
21	84M726	KIPP STAR COLLEGE PREP CHARTER SCHOOL

It should be noted that in given data that contains information about SHSAT registration (`shsat_data`), a school may have multiple rows of data because it takes into account the year. However, the other data set (`school_data`) has just one row for each school and does not specify the year.

Possible ways to take this into account:

- Try creating models for each year
- Take average over the years for each school

Either way, we'd only have 21 training points but 1251 testing points.

5.2 Linear Regression

5.3 Random Forest and Variance Importance

6 Conclusions

Lorem ipsum dolor sit amet, an eos tation consequuntur, vis bonorum mediocritatem cu. Mel erat legere id. Vis ei agam omnesque, in pri reque volutpat conceptam, ad nihil timeam lucilius nec. Sed ne verterem tacimates. Tritani scaevola nec ex, sint doming tacimates ei mea. Nam ea blandit invidunt. Vix ne nusquam placerat democritum. Nam ei vidisse debitis, at malis doming sed. Ea dicit efficiendi pro. Iudico iisque accommodare ei per.

Eam quot delicata ut. Natum nusquam definitiones ei qui. Enim cetero euismod cu usu, noster luptatum ea vis. Ex deserunt maiestatis sit, et saepe vidisse appareat vix. Vidit adhuc has in, his suscipit mediocritatem ex, mei prima suscipiantur an. Quas regione adversarium has ei, falli appareat voluptaria vel ei.