Introduction to Data Science – DS UA 112

Final project

In this final project, we will explore a dataset provided by the New York City Department of Education. Of particular interest is whether characteristics of NYC middle schools predict admission to one of 8 highly selective public high schools (Stuyvesant, Bronx High School of Science, etc.) in New York (from now on called HSPHS). Admission to these schools is contingent on applying AND scoring sufficiently highly on the Specialized High Schools Admissions Test (SHSAT), an independently produced and anonymously graded standardized test.

The Dataset: The dataset ('middleSchoolData.csv') contains data from all 594 NYC middle schools, including 485 public schools and 109 charter schools (in the last 109 rows) from a randomly picked year in the past 5 years. Each row of the dataset represents a particular school, so the unit of analysis is "school".
Here is what the columns represent:
A - B: NYC DOE school code and name, respectively
C: Number of applications to HSPHS originating from this school
D: Number of applicants to HSPHS accepted from this school
E: Per student spending, in $
F: Average class size
G-K: Self-described ethnic identity of the student body
L-Q: Average rating of "school climate" factors as perceived by the students, e.g. trust
R: Percentage of students who have been evaluated as disabled
S: Percentage of students living in households below the poverty line
T: Percentage of ESL students
U: School size (Number of students in the entire school)
V: Average student achievement on a state-wide standardized test
W-X: Proportion of students exceeding state-wide expectations in reading and math

This dataset is comprehensive, but some data is missing. If data in a cell is missing, you have to handle (clean or impute) it, in order to do the analyses. Sometimes, data is missing systematically. For instance, data for columns E and F is missing for all charter schools.

Format: The project is comprised of your answers to 10 questions. Each answer should ideally include some paragraph of text (describing what you did and what you found), a figure that illustrates the findings and some numbers (e.g. test statistics or p-values). Please save it as a word, pdf or pages document. This document should be 4-6 pages long (arbitrary font size and margins). ~half a page per question is reasonable. In addition, open your document with a brief statement as to how you handled dimension reduction, data cleaning and data transformation, as this will apply to all answers. Make sure to include your name.

Academic integrity: You are expected to do this project by yourself, individually so that we are able to determine a grade for you. There are enough degrees of freedom (e.g. how to clean the data, what variables to compare, aesthetic choices in the figures, etc.) that no two reports will be alike. We'll be on the lookout for suspicious similarities, so please refrain from collaborating.

Questions:

1) What is the correlation between the number of applications and admissions to HSPHS?
2) What is a better predictor of admission to HSPHS? Raw number of applications or application *rate*?
3) Which school has the best *per student* odds of sending someone to HSPHS?
4) Is there a relationship between how students perceive their school (as reported in columns L-Q) and how the school performs on objective measures of achievement (as noted in columns V-X).
5) Test a hypothesis of your choice as to which kind of school (e.g. small schools vs. large schools or charter schools vs. not (or any other classification, such as rich vs. poor school)) performs differently than another kind either on some dependent measure, e.g. objective measures of achievement or admission to HSPHS (pick one).
6) Is there any evidence that the availability of material resources (e.g. per student spending or class size) impacts objective measures of achievement or admission to HSPHS?
7) What proportion of schools accounts for 90% of all students accepted to HSPHS?
8) Build a model of your choice – clustering, classification or prediction – that includes all factors – as to what school characteristics are most important in terms of a) sending students to HSPHS, b) achieving high scores on objective measures of achievement?
9) Write an overall summary of your findings – what school characteristics seem to be most relevant in determining acceptance of their students to HSPHS?
10) Imagine that you are working for the New York City Department of Education as a data scientist (like one of my former students). What actionable recommendations would you make on how to improve schools so that they a) send more students to HSPHS and b) improve objective measures or achievement.

Hints:

*In order to do some analyses, you might have to apply a dimension reduction method first. Key candidates for such methods are the 6 school climate variables and the 3 objective achievement variables. Can you express them with fewer independent factors that capture the same information?
*In order to do some analyses, you will have to clean the data first, either by removing or imputing missing data (either is fine, but explain and justify what you did)
*If you encounter highly skewed data, you might want to transform the data before doing anything, e.g. z-scoring, using log-transforms or the like.
*For questions 1 and 6, seeing a scatter plot would be nice, to illustrate the correlations.
*To convert numbers to rates (as in question 2), divide numbers by school size
*For question 5 – the hypothesis test – you might have to transform some variables into categorical variables first, e.g. converting student spending into “above and below the median” to convert to “rich” vs. “poor” schools. Don’t restrict yourself to that comparison. You could apply this logic to almost every variable to – then – use it as an independent variable here.
*For question 7, a bar graph of schools, rank-ordered by decreasing number of acceptances of students to HSPHS would be nice to see.
*For questions 9 and 10, no figures are necessarily needed – a narrative (based on the answers to the other questions) is sufficient.