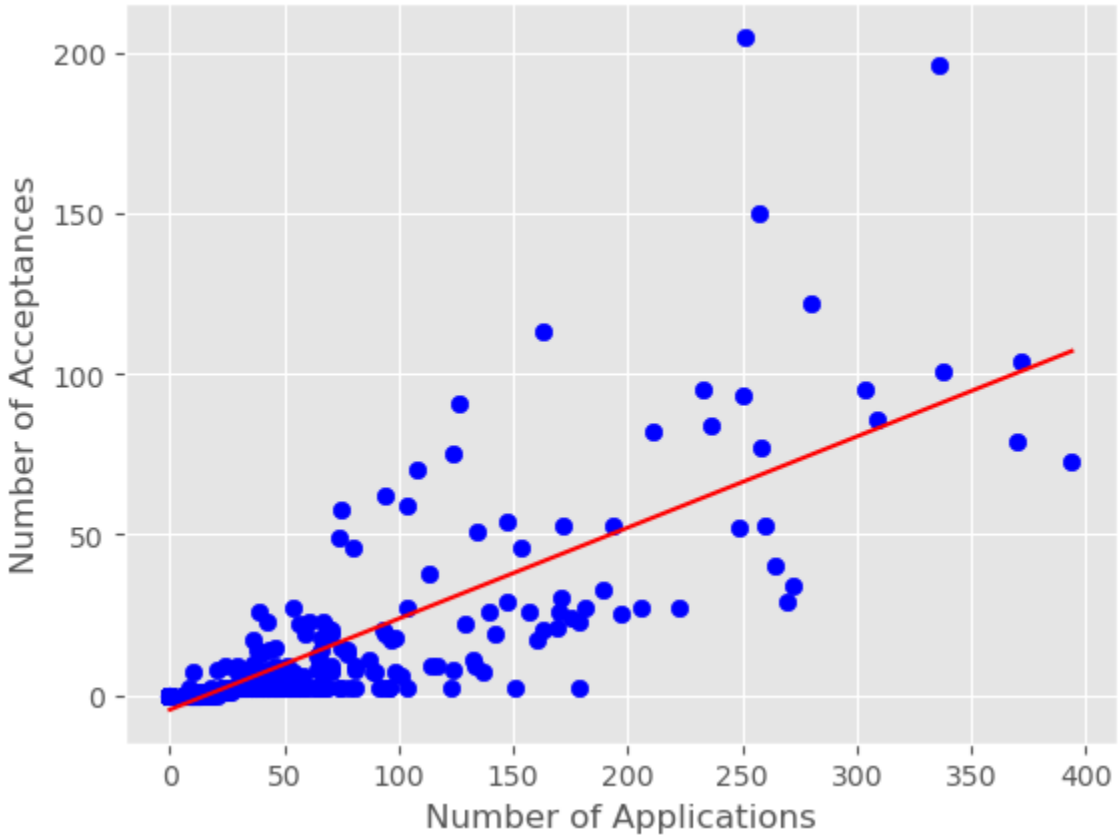
Introduction to Data Science – DS UA 112

Final project

Maryam Khalili (mk6323)

---

1) What is the correlation between the number of applications and admissions to HSPHS?



There is a high, positive correlation of 0.8 between number of applications and admission to HSPHS. This correlation is not really meaningful because obviously that there will be more acceptances among a larger pool of applications than a smaller one. From this graph we can also calculate the acceptance rate of a school with a particular number of applications. Acceptance rate is tricky and there are a lot of factors that need to be considered. That's why I think it's not really meaningful when schools pride themselves on their acceptance rate. When a school is extremely prestigious it receives less number of applicants and thus the acceptance rate is not

really high, in contrast to a school where everybody thinks they have a shot at and this brings down the acceptance rate by a lot.

2) What is a better predictor of admission to HSPHS? Raw number of applications or application *rate*?

{Application rate = number of applications / school size}

Intuitively application rate would be a better predictor than application numbers. To understand this better I did a linear regression between number of acceptances and number of applications and also number of acceptances and applications rate. Below are the summaries of the two regression models:

regression between acceptances and number of applications:

OLS Regression Results

Dep. Variable:	acceptances	R-squared (uncentered):	0.655
Model:	OLS	Adj. R-squared (uncentered):	0.655
Method:	Least Squares	F-statistic:	1126.
Date:	Tue, 18 May 2021	Prob (F-statistic):	3.46e-139
Time:	22:51:27	Log-Likelihood:	-2374.1
No. Observations:	594	AIC:	4750.
Df Residuals:	593	BIC:	4755.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
applications	0.2472	0.007	33.560	0.000	0.233	0.262

Omnibus:	670.446	Durbin-Watson:	1.758
Prob(Omnibus):	0.000	Jarque-Bera (JB):	56546.871
Skew:	5.245	Prob(JB):	0.00
Kurtosis:	49.634	Cond. No.	1.00

regression between acceptances and application rate:

OLS Regression Results

Dep. Variable:	acceptances	R-squared (uncentered):	0.440
Model:	OLS	Adj. R-squared (uncentered):	0.439
Method:	Least Squares	F-statistic:	464.8
Date:	Tue, 18 May 2021	Prob (F-statistic):	1.68e-76
Time:	22:51:27	Log-Likelihood:	-2509.8
No. Observations:	592	AIC:	5022.
Df Residuals:	591	BIC:	5026.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	177.6172	8.238	21.560	0.000	161.437	193.797

Omnibus:	601.434	Durbin-Watson:	1.741
Prob(Omnibus):	0.000	Jarque-Bera (JB):	29077.147
Skew:	4.594	Prob(JB):	0.00
Kurtosis:	36.082	Cond. No.	1.00

But after comparing the regression models we see that the number of applications is a better predictor since it has a bigger R-squared (0.655) and smaller std_err (0.007). Therefore we conclude that in this specific case, the number of applications is a better predictor of acceptance to HSPHS, as we already saw in the correlation in the previous question.

3) Which school has the best *per student* odds of sending someone to HSPHS?

{Per student rate = number of acceptances/school size}
The Christa Mcauliffe School\I.S. 187 (index 304) with a rate of 0.235 has the best per student rate among the 594 schools.

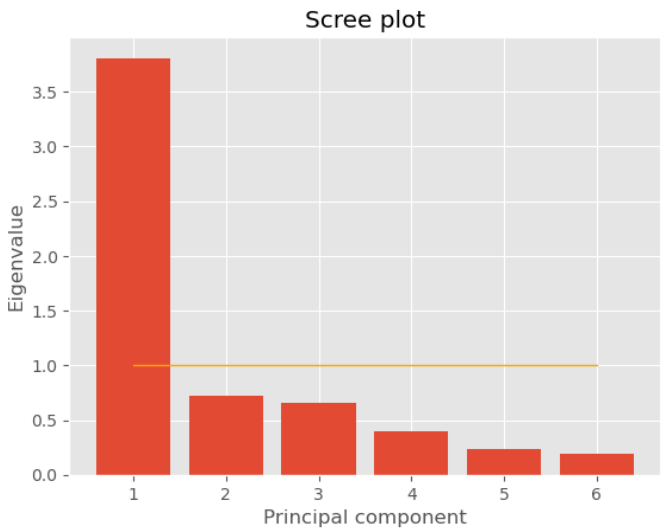
4) Is there a relationship between how students perceive their school (as reported in columns L-Q) and how the school performs on objective measures of achievement (as noted in columns V-X).

Group 1 - School features
L-Q: rigorous_instruction, collaborative_teachers, supportive environment, effective school leadership, strong family-community ties, trust

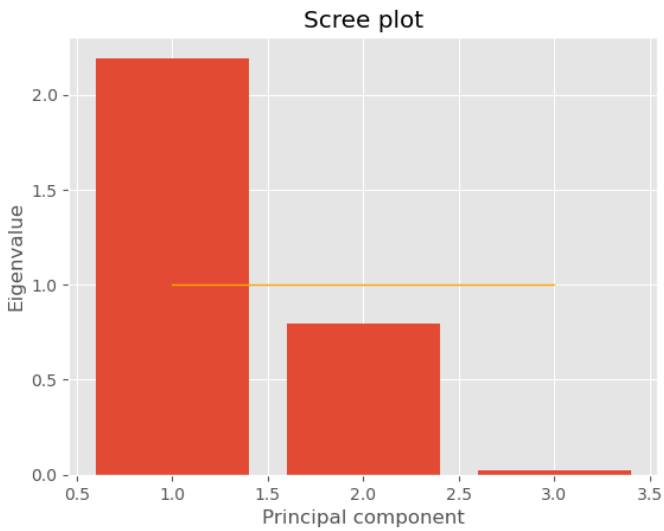
Group 2 - Student success
V: Average student achievement on a state-wide standardized test
W-X: Proportion of students exceeding state-wide expectations in reading and math

To find out if there is a relationship we will do a Pearson correlation. But before that we have to do a dimension reduction first because in both groups we can reduce the factors to a single meaningful one. We are going to run PCA (Principal Component Analysis) for both groups and below are the results:

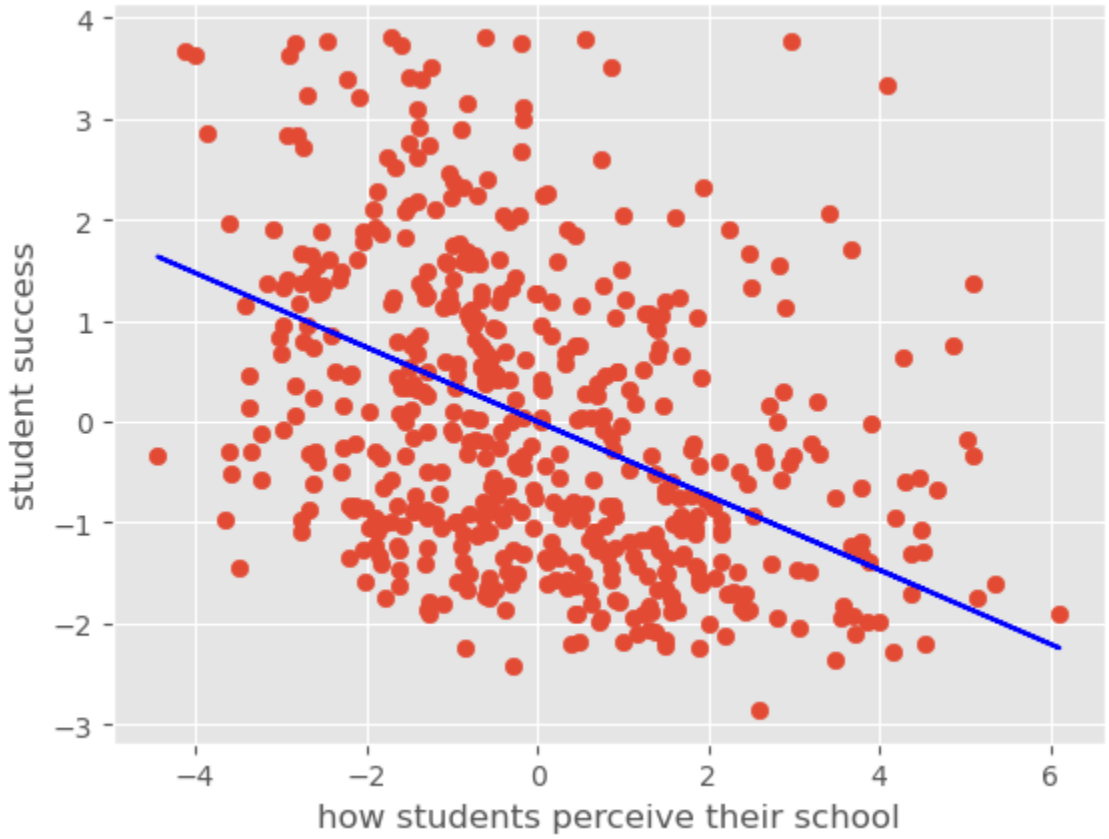
Group 1:



Group 2:



Group 1 or school features can basically be interpreted as ‘how good students think their school is’ and group 2 can be interpreted as ‘how good students perform’. After correlating the new components we get a (very surprising) correlation of -0.36. Meaning the better a student perceives their school the less successful they are (!):



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).

I decided to test my hypothesis on whether public schools versus charter schools perform differently on sending students to HSPHS (number of admissions). I chose to do a t-test because it's an effective way to study whether two parametric groups are different from one another and how significant this difference is.

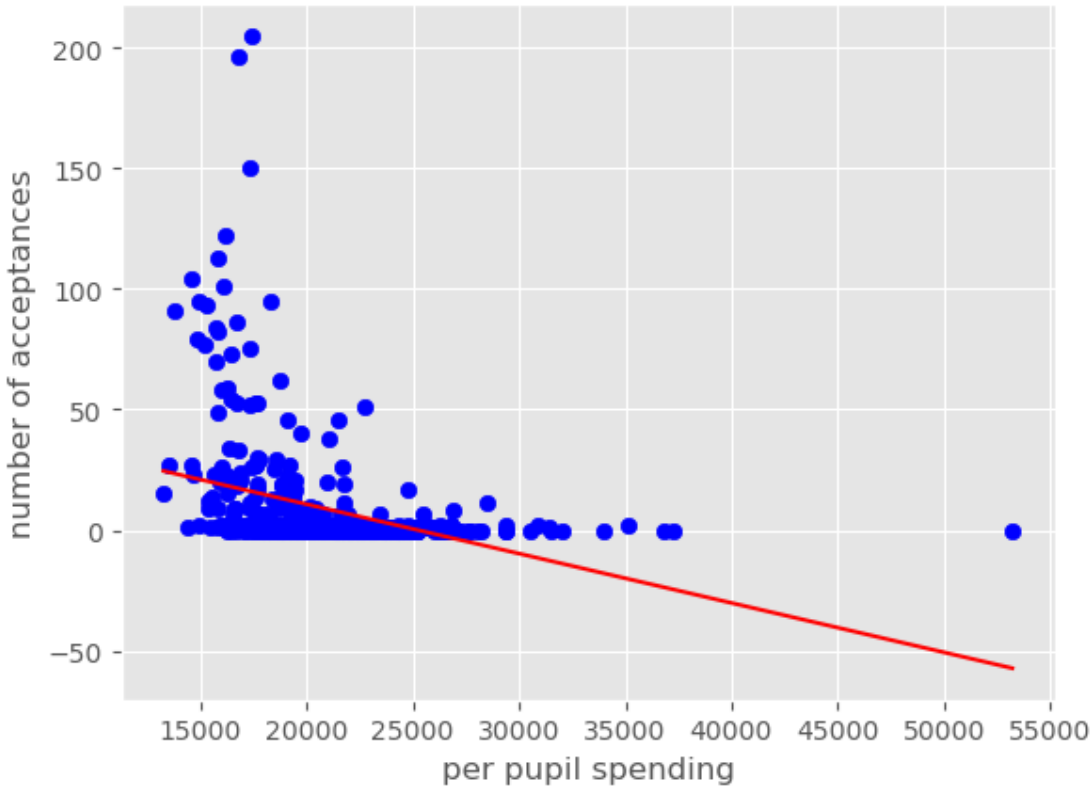
Null hypothesis: there is no difference -that could not have happened by chance- between the number of students that public schools and charter schools send to HSPHS.
P-value =~ 0.001
Conclusion: We conclude that this difference could not be explained by chance alone and thus reject the null hypothesis for an alpha < 0.05. Of course this is not enough to conclude that one group of school is performing better solely based on the number of acceptances. Public schools have a bigger school size and a larger pool of applicants which increases the number of acceptances (as we saw in the first question).

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?

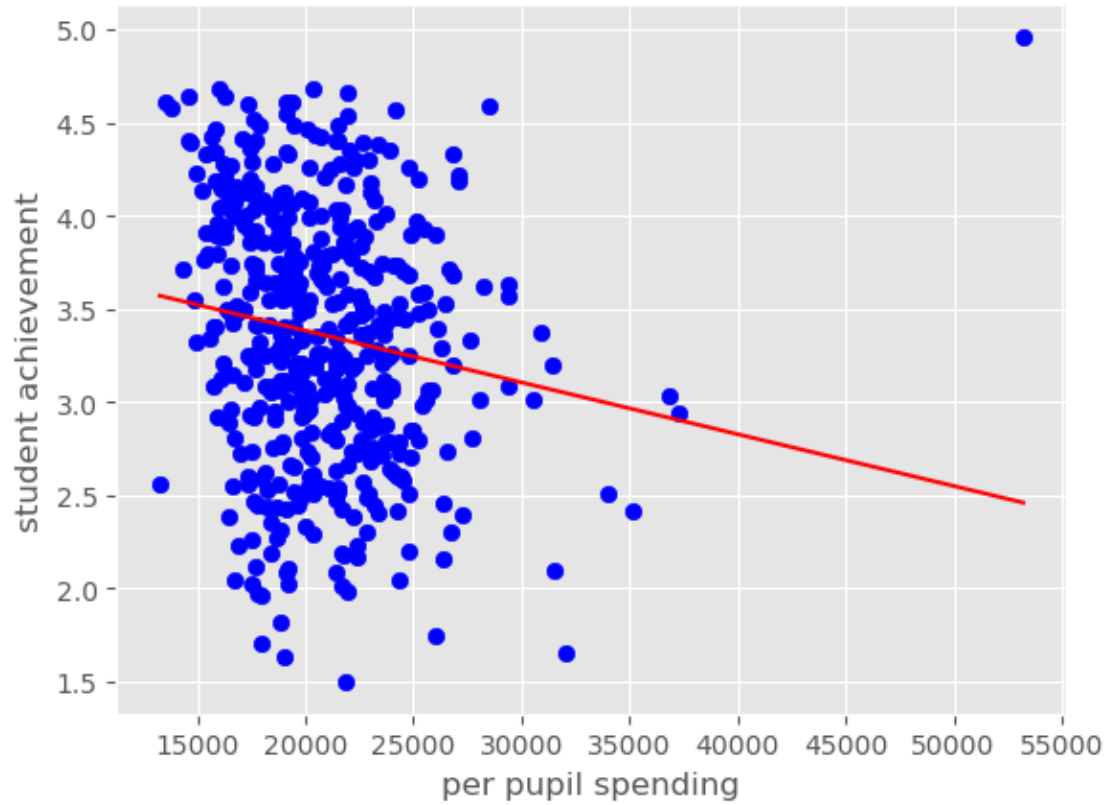
To answer this question I did four correlations:

-- continued next page

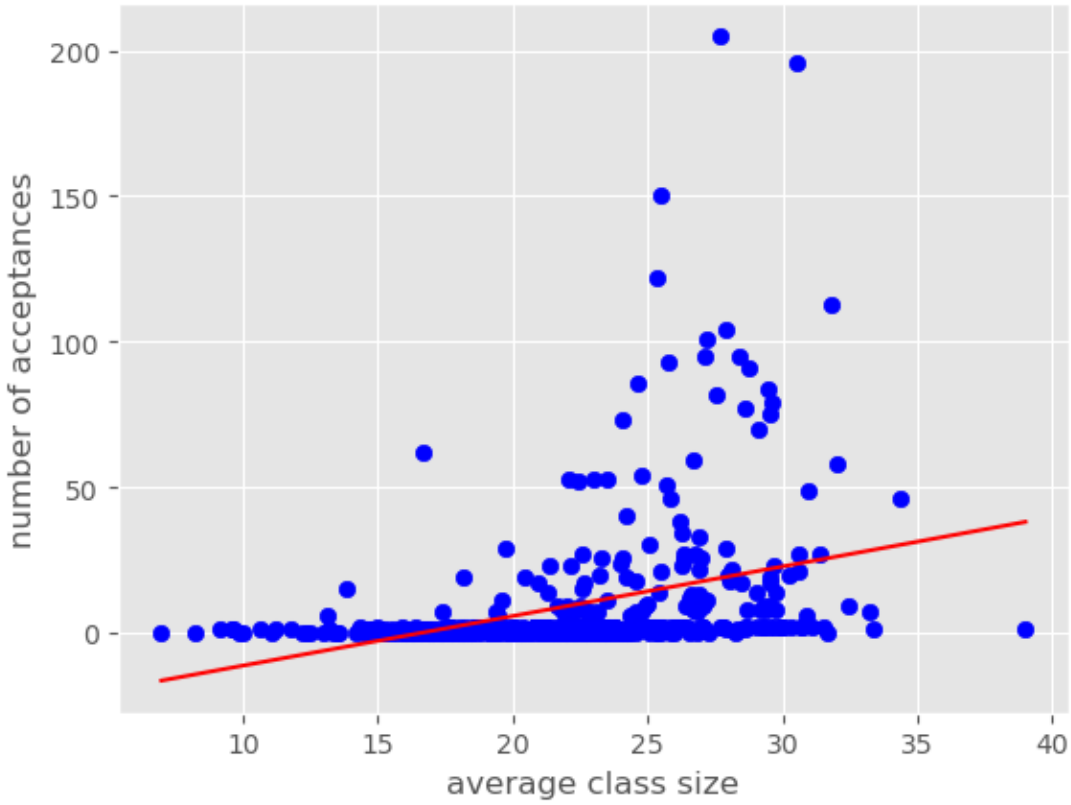
1) There is a correlation of -0.33 between per pupil spending and number of acceptances.



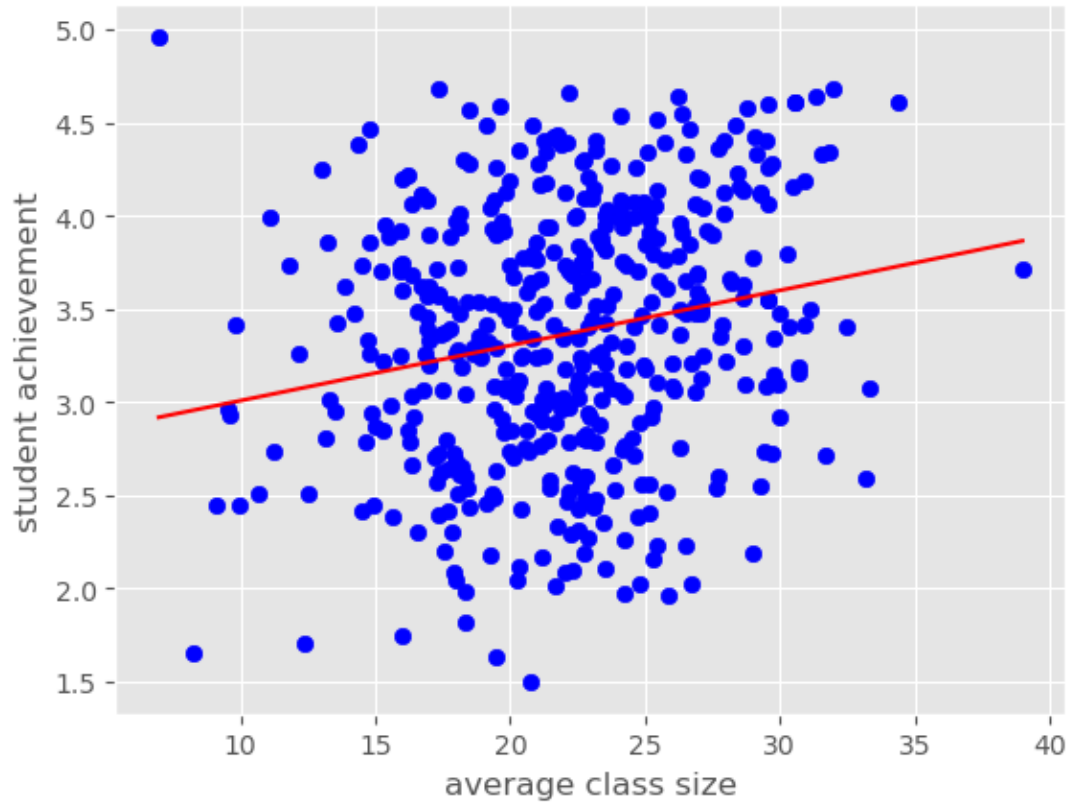
2) There is a correlation of -0.15 between per pupil spending and student achievement.



3) There is a correlation of 0.34 between average class size and number of acceptances.



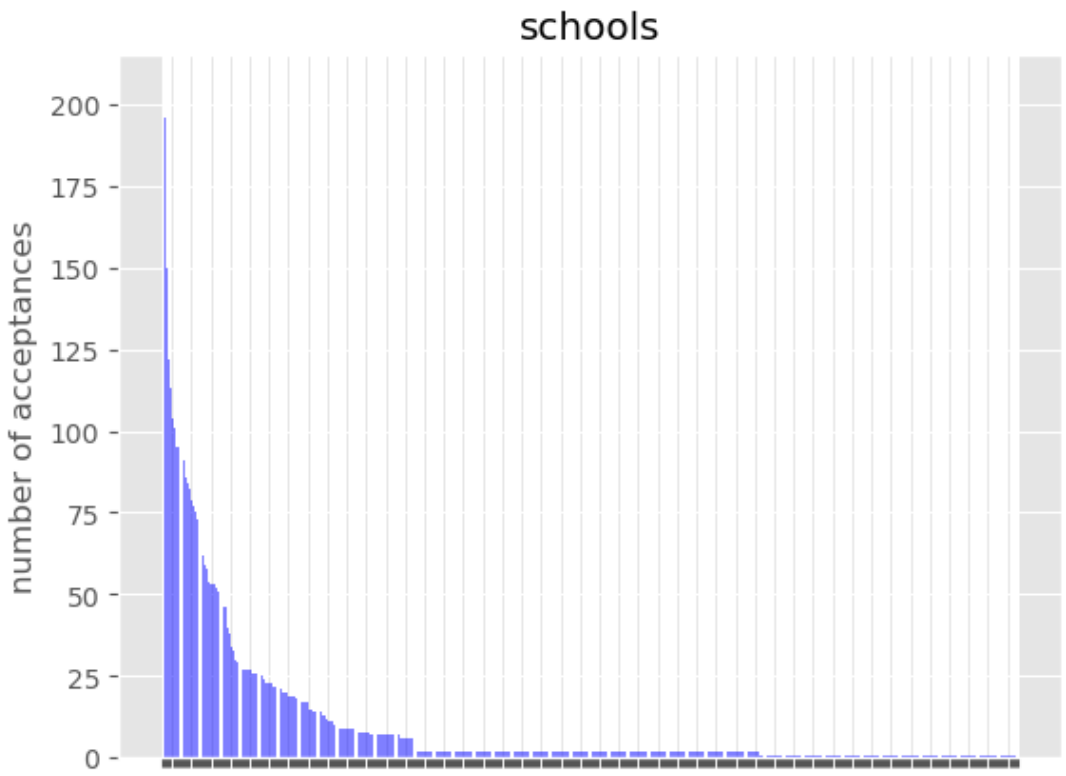
4) There is a correlation of 0.2 between average class size and student achievement.



7) What proportion of schools accounts for 90% of all students accepted to HSPHS?

Rephrasing this question to make calculations easier:
Total number of acceptances is 4461. 90% of this population is 4015. How many schools contribute to this number? (407 schools)

```
count = 0
for i in reversed(sorted_acceptances):
    while count <= 4015:
        count += sorted_acceptances[i]
        i += 1
        # print(i)

# we find out there are 407 schools that account for 90% of the total acceptances

# we make a dictionary with all the schools and their acceptance #, sort them based
on their
# value, keep the values from index 0 to 407 and then plot them

schools = data["school_name"]
acceptances = data["acceptances"]

d = {}
for A, B in zip(schools, acceptances):
    d[A] = B

d = sorted([(v, k) for k, v in d.items()], reverse=True)

# wait there are actually 407? idk
ninety_percent = d[:407]
print(ninety_percent)

schools = np.array([])
acceptances = np.array([])

for i in range(len(ninety_percent)):
    schools = np.append(schools, ninety_percent[i][1])
    acceptances = np.append(acceptances, ninety_percent[i][0])
```

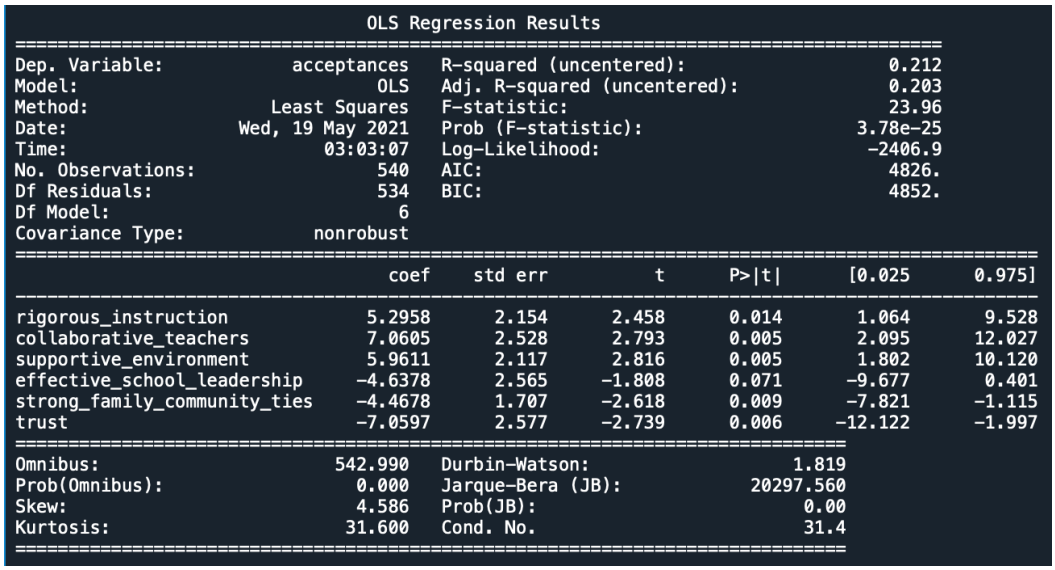


Schools shown in this bar chart account for 90% of the total number of acceptances to HSPHS.

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?

I chose to build a prediction model using multiple regression. Multiple regression is useful here because it helps us predict the value of a variable based on the value of other variables. In our case, we are going to see which of our variables is the most important in terms of student acceptance to HSPHS and student success.

1) Multiple regression between school characteristics (L-Q) and number of acceptances:



OLS Regression Results

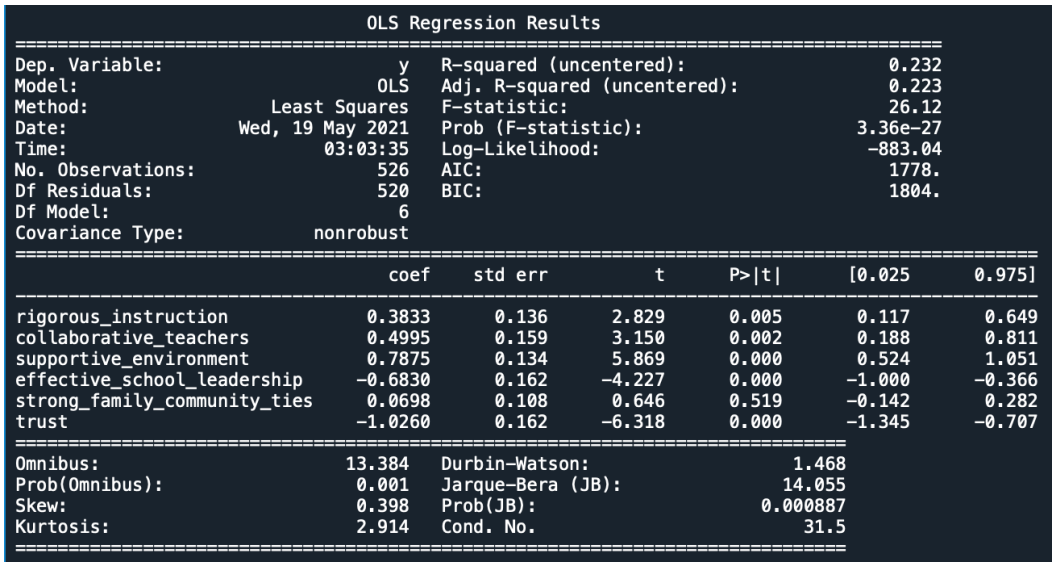
Dep. Variable:	acceptances	R-squared (uncentered):	0.212
Model:	OLS	Adj. R-squared (uncentered):	0.203
Method:	Least Squares	F-statistic:	23.96
Date:	Wed, 19 May 2021	Prob (F-statistic):	3.78e-25
Time:	03:03:07	Log-Likelihood:	-2406.9
No. Observations:	540	AIC:	4826.
Df Residuals:	534	BIC:	4852.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
rigorous_instruction	5.2958	2.154	2.458	0.014	1.064	9.528
collaborative_teachers	7.0605	2.528	2.793	0.005	2.095	12.027
supportive_environment	5.9611	2.117	2.816	0.005	1.802	10.120
effective_school_leadership	-4.6378	2.565	-1.808	0.071	-9.677	0.401
strong_family_community_ties	-4.4678	1.707	-2.618	0.009	-7.821	-1.115
trust	-7.0597	2.577	-2.739	0.006	-12.122	-1.997

Omnibus:	542.990	Durbin-Watson:	1.819
Prob(Omnibus):	0.000	Jarque-Bera (JB):	20297.560
Skew:	4.586	Prob(JB):	0.00
Kurtosis:	31.600	Cond. No.	31.4

R-squared or the coefficient of determination is the proportion of variability in y (in this case acceptances) that can be explained by x (school features). So in our case, 21% of the acceptances is due to school features like collaborative teachers, supportive environment, rigorous instruction etc.

2) Multiple regression between school characteristics (L-Q) and achieving high scores on objective measures of achievement (V-X) (after doing PCA):



OLS Regression Results

Dep. Variable:	y	R-squared (uncentered):	0.232
Model:	OLS	Adj. R-squared (uncentered):	0.223
Method:	Least Squares	F-statistic:	26.12
Date:	Wed, 19 May 2021	Prob (F-statistic):	3.36e-27
Time:	03:03:35	Log-Likelihood:	-883.04
No. Observations:	526	AIC:	1778.
Df Residuals:	520	BIC:	1804.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
rigorous_instruction	0.3833	0.136	2.829	0.005	0.117	0.649
collaborative_teachers	0.4995	0.159	3.150	0.002	0.188	0.811
supportive_environment	0.7875	0.134	5.869	0.000	0.524	1.051
effective_school_leadership	-0.6830	0.162	-4.227	0.000	-1.000	-0.366
strong_family_community_ties	0.0698	0.108	0.646	0.519	-0.142	0.282
trust	-1.0260	0.162	-6.318	0.000	-1.345	-0.707

Omnibus:	13.384	Durbin-Watson:	1.468
Prob(Omnibus):	0.001	Jarque-Bera (JB):	14.055
Skew:	0.398	Prob(JB):	0.000887
Kurtosis:	2.914	Cond. No.	31.5

R-squared explains that 23% of the student achievement is due to school features mentioned above.

NOTE: By student success/objective measures of achievement I'm referring to the last three columns:
V: Average student achievement on a state-wide standardized test
W-X: Proportion of students exceeding state-wide expectations in reading and math

---

9) Write an overall summary of your findings – what school characteristics seem to be most relevant in determining acceptance of their students to HSPHS?

Collaborative teachers (coef = 7), supportive environment (coef = 6), and rigorous instruction (coef = 5.3) with high coefficients and high t stats (which measures how statistically significant the coefficient is) are the most important factors that contribute to student admission to HSPHS. Factors like trust and effective school leadership are the least important factors in the students' admission. (based on the regression model in question 8)
The same factors - supportive environment (coef = 0.8) and collaborative teachers (coef = 0.5) with high t stats and low standard error are the most important in determining the objective measures of achievement of students. The least important factors for student achievement are trust and effective school leadership.

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.

Based on my findings, having a supportive environment and collaborative teachers are very important in student acceptances to HSPHS and also their other achievements. This is especially important when a school has a big student population. Since most applications come from

schools with a large number of students, it is vital that they are provided with enough support and helpful teachers. So I would recommend a series of training for teachers and administrators on how to be more helpful and effective. This will also help with having a more supportive environment. Helpful teachers will also encourage students to be more compassionate and collaborative with one another. Having a more supportive environment does not necessarily imply spending more money per student since it has not proven very effective in neither admission nor student success. There should be other forms of encouragement and awarding students that are not monetary.

Another important finding was that average class is positively correlated with both acceptance to HSPHS and student achievement. This might be because more students get motivated to as they see others who work hard for their applications to HSPHS. Positive influence is really important and effective especially in a large crowd.

I would recommend increasing average class size, having/training more collaborative teachers and building a supportive environment to have more students admitted to HSPHS and succeed beyond expectations.