Alan Chu

Final Project

0) How did you handle dimension reduction, data cleaning and data transformation?

```
np.where(pd.isnull(df["column_name"]))
```

I used the code above to check which cells are null in the columns and to determine if data cleaning is necessary.

```
df = df.dropna(subset=["column_name"])
```

The code above is used to remove rows with null elements within the specified columns.

```
df=(df-df.mean())/df.std()
```

The code above is used to normalize data

```
fill_NaN = SimpleImputer(missing_values=np.nan, strategy='mean')
df = pd.DataFrame(fill_NaN.fit_transform(df))
```

The code above is used to impute data using the mean strategy

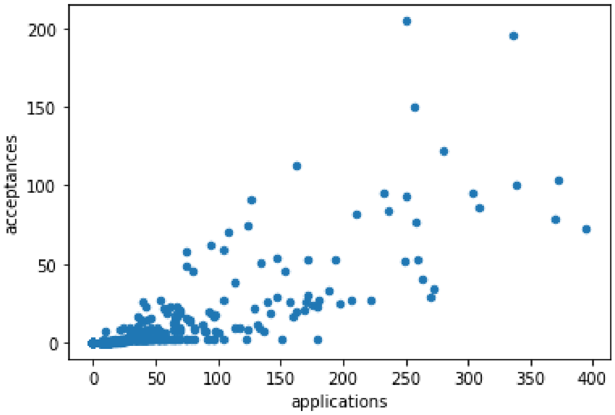
1) What is the correlation between the number of applications and admissions to HSPHS?

*data cleaning was not necessary because there were no empty cells in the application and acceptances columns.

```
x = df["applications"]
y = df["acceptances"]
correlation = x.corr(y)
print(correlation)
```

Using pandas correlation function, I found that the correlation between applications and acceptances is .8017. The two variables have a high positive correlation which suggests that as the number of applications increases, the number of acceptances will increase as well. The scatter plot displays a positive correlation as well.

```
df.plot.scatter("applications","acceptances")
```



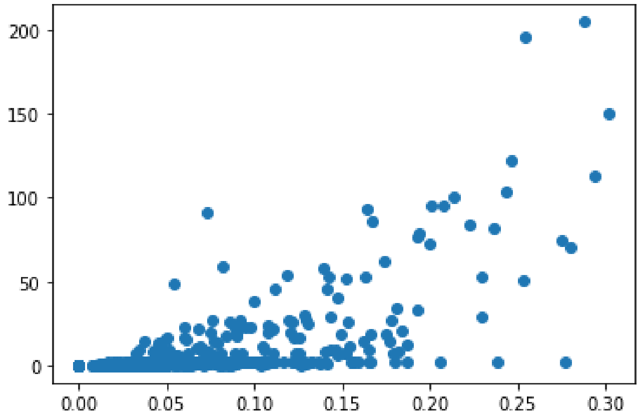
2) What is a better predictor of admission to HSPHS? Raw number of applications or application *rate*?

*imputer missing values in school size

```
applicationRates = df["applications"] / df["school_size"]
```

To calculate the application rates of each school, I divided the number of applications by the student population. After normalizing the data, I found that the correlation coefficient between the application rate and the acceptances is .6588. This is lower than the correlation coefficient between number of applicants and acceptance of .8017 which was calculated in part 1. Because the raw number of applicants produced a higher correlation coefficient after normalization of the data, it would be a better predictor of admission. Furthermore, I used a simple linear regression for both variables in which applications had an r-squared value of .65 while application rate had a r-squared value of .47. This indicates that the 65% of the data fit the regression model of number of applications, making more likely to be a better predictor

Applications rate vs Acceptances graph



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

*no data cleaning needed because acceptances and applications did not have any null values

```
acceptanceRates  = df["acceptances"] / df["applications"]
maxVal = np.argmax(acceptanceRates)
```

To find the odds of a student being accepted into HSPS, I calculated the acceptance rate of each student by dividing the number of acceptances by the number of applications. The group of applicants serve as the sample for the school population. Using the argmax function, I found the element ID to print the name of the school with the highest admission rate. “THE CHRISTA MCAULIFFE SCHOOL\I.S. 187” has the best “per student” odds of sending someone to HSPHS of 81.67% chance.

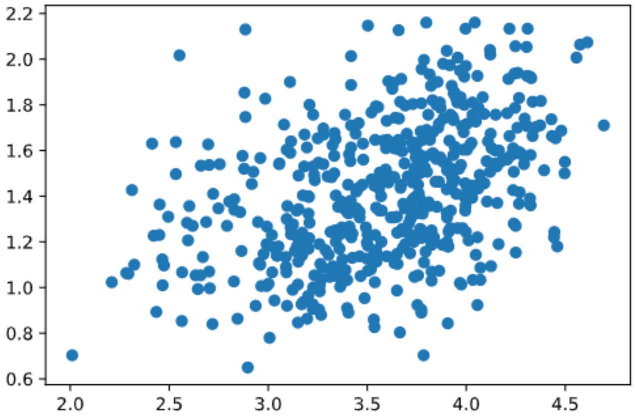
4) Is there a relationship between how students perceive their school (as reported in columns L-Q) and how the school performs on objective measures of achievement (as noted in columns V-X)

* removed rows with null values within columns L-Q and V-X

```
df=(df-df.mean())/df.std()
x = (df["rigorous_instruction"] +...+df["trust"])/6 #adds columns L-Q
y = (df["student_achievement"] + df["reading_scores_exceed"] + df["math_scores_exceed"])/3
correlation = x.corr(y)
```

To measure how the students perceive their school, I combined the multiple ratings and averaged it for an overall perception rating. Since a high rating in each column could be interpreted as “good” and vice versa, the combined score would essentially be rating from “bad” to “good” perception of the school. The same logic could be applied for the columns of the objective measures. Because “student_achievement” and “math/reading_scores _exceeded” are on different scales, I normalized the data and then found the average for an overall objective measure of achievement. The correlation between the perception and achievement was .3969 which suggests that there is a moderate positive correlation between the students perception of the school and academic performance.

Overall Perception Score vs Overall Achievements Score



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).

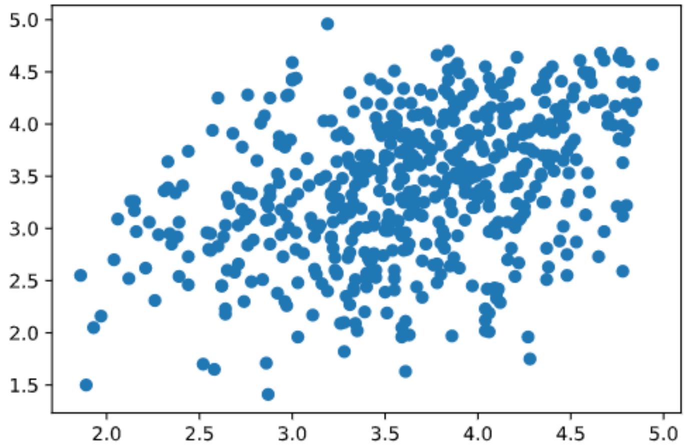
*dropped null values in rigorous instruction and student achievement

```
for x in range(len(RI)):
    if(RI[x] <= ave):lowRI.append(SA[x])
        else: highRI.append(SA[x])
    t = stats.ttest_ind(lowRI,highRI)
```

```
Output: statistic=-8.335303140374153, pvalue=6.7066787076887e-16
```

I hypothesize that schools with high rigorous instruction ratings will yield higher student achievement scores because strict rules may encourage students to pay more attention to their studies. To determine which schools had high or low rigorous rating scores, schools with student achievement scores lower than the overall average were considered as low and vice versa. Then I conducted a hypothesis test. The null hypothesis is that the mean of scores for low rigorous ratified schools is equal to the mean of scores for high rigorous rated schools. The alternative hypothesis is that the two means would not be equal. Since the test statistic was equal to 8.335 and higher than z(.05) = 1.96, we can reject the null hypothesis at a significance level of .05. Further, the p value of near 0 was less than .05, there is more evidence to reject the null hypothesis and accept the alternative hypothesis. High rigorous instruction is likely to result in higher student achievement scores.

Rigorous Instruction Rating vs Student Achievement Score



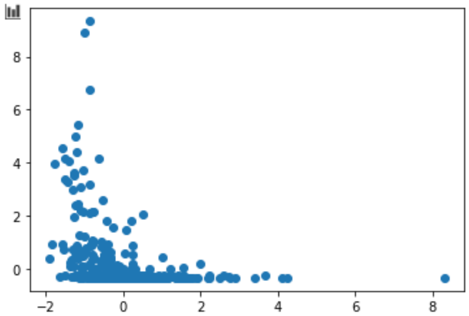
6) Is there any evidence that the availability of material resources (e.g. per student spending or class size) impacts objective measures of achievement or admission to HSPHS?

*dropped null values which includes all charter schools

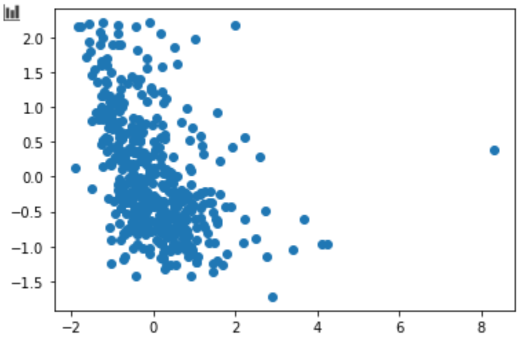
```
print(correlations)
Output:
      perPupilSpending vs Acceptances:-0.3369726734526743
 perPupilSpending vs Achievement Score:-0.44290689081093737
avgClassSize vs Achievement Acceptances: 0.34868632026641133
     avgClassSize vs Achievement Score: 0.5048063052192899
```

Because charter schools did not have values for the material resources variables, the information only applies to non-charter schools. Availability of material resources appears to have an impact on both measures of achievement and admissions to HSPHS because each has moderate correlation value being within the .3 to .5 range. Per student spending appeared to have a negative correlation with both academic measures and number of admissions. This indicates schools with lower funding will more likely have a higher number of acceptances for HSPHS and higher academic scores. Although the idea of less funding may seem controversial, it may just mean the funding does not impact the academic results. In addition, average class size had a positive correlation with both dependent variables.This indicates that schools with larger class sizes will more likely yield higher acceptances and academic achievement. Based on the correlation values, objective measures of academic achievement is more affected by the material resources than the number of students accepted to HSPHS.

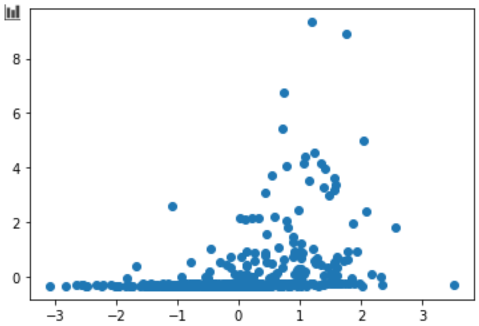
perPupilSpending vs Acceptance



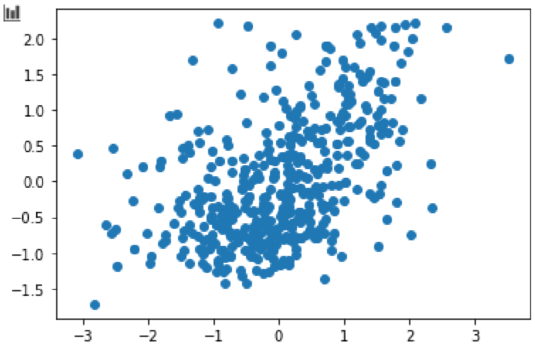
perPupilSpending vs Achievement Score



avgClassSize vs Achievement Acceptances



avgClassSize vs Achievement Score



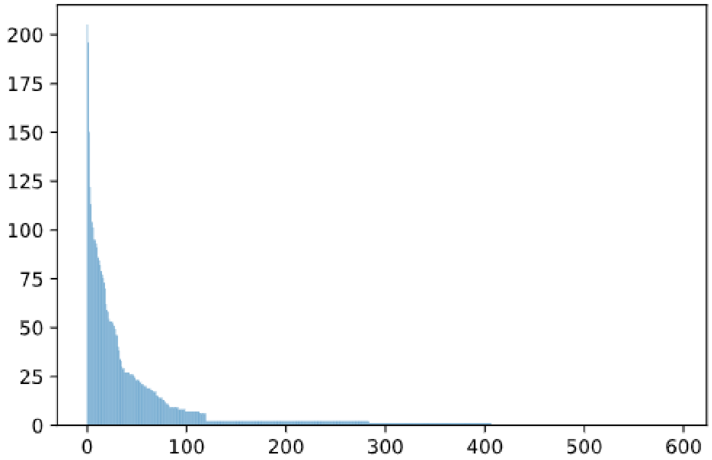
7) What proportion of schools accounts for 90% of all students accepted to HSPHS?

*no empty cells in acceptances column

```
acceptances = np.sort(acceptances)[::-1]
x = np.sum(acceptances) * .9
for i in acceptances:
    if(total >= x): break
    else:
        total += i
        count += 1
proportion  = count/len(acceptances)
```

The variable x is equal to 90% of all students accepted to HSPHS., I added the number of acceptances from each school in descending order until the total number exceeded the 90% mark while keeping track of the number of schools. By dividing the number of schools by the total number of schools, I found that the proportion of schools that accounts for 90% of all acceptances is equal to **20.7%**.

Bar Graph of Acceptances for each school



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?

*dropped null values

X = sm.add_constant(1)
result = sm.OLS(y2, X).fit()

To determine which independent variables were significant, I performed a multiple regression on each factor on the number of admissions and calculated the p-values. In the initial regression model,the factors applications, school size, and poverty percentage had the lowest p values. R-squared value was equal to .704 which indicated 70.4% of the data fits themultiple linear regression equation.
*p-value for poverty percentage was slightly higher than .05

Acceptances = 21.06 + 0.3205(applications) + -0.0136(school_size) + -0.2384(poverty_percent)

I applied the same logic to find a multiple linear regression model on objective measures of achievement. To calculate achievement, I normalized the student achievement score, math score exceeded, and reading score exceeded and then averaged them.R-squared value was equal to .669 which indicates that 66.9% of the data fits the multiple linear regression equation.

Overall academic score = 0.6018 + 0.0024(applications) + 0.3679(rigorous rating)
+ -0.0237(disability percentage) + -0.0199(poverty percentage)

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

There is a positive correlation between the raw number of applications and acceptances. The application rate is not as good of a predictor as raw number of applications because it has a lower correlation and r-squared value. The school with the best"per student" odds of sending someone to HSPHS is The Christa Mcauliffe School. There is a positive correlation between how students perceive the school and their academic performance. Schools with higher rigorous instructions are more likely to have a higher academic achievement score. The overall objective measure of academic success is influenced by the material resource factors than the number of students accepted to HSPHS. 20.7% of schools could account for 90% of all acceptances to HSPHS. Overall, the best school characteristics in determining acceptance of their students to HSPHS are school size, number of applications, and poverty percentage. They all have a low p value and a high r-squared value when used in a multiple linear regression model.

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.

a) To increase the amount of acceptances to HSPHS, I would recommend schools decrease their average class size or hire more teachers so that teachers are able to put more focus on each individual student. Schools should also encourage students to apply to HSPHS which will at least give them a chance to be accepted. Offering funds to the students in poverty may assist their studies and increase their chances of getting into HSPHS.

b) To improve objective measures of achievement, I would recommend enforcing a more rigorous education style so that students will be more inclined to study and raise their grades. Providing some form of reward for getting accepted to HSPHS may encourage students to be more serious about applying and willing to improve their academics to get accepted. Schools should also put more focus on the students who are disabled or in poverty because typically they would need more educational assistance

-Thank you for a great semester. I learned a lot and had fun while doing so!
Sincerely,
Alan Chu