Evan LaBanca

Final Project

To read in the csv file, I ran

```
df = pd.read_csv('middleSchoolData.csv').
```

To delete the rows that contained null values, I would run the following command. I would do this on the columns that I would be working with for each question.

```
df = df.dropna(how='any', subset=[])
```

To normalize the data, I would run the following command. I would do this when the data that I'm working with are not of the same scale. This is z-scoring.

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

I would also be importing packages such as pandas, numpy, matplotlib.pyplot, and scipy.stats.

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

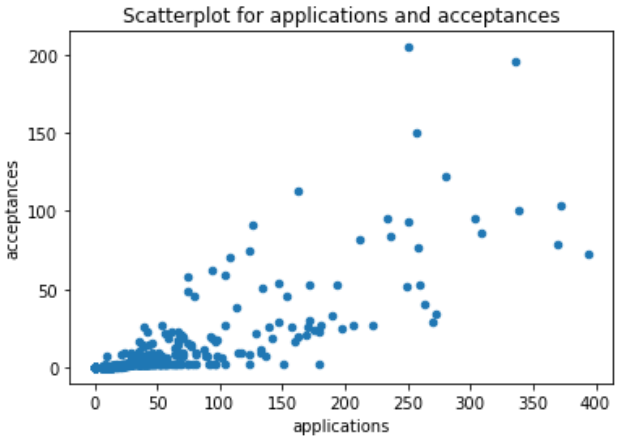
To begin, I began by running

```
df = df.dropna(how='any', subset=['applications', 'acceptances'])
```

I would then run the following command

```
correlation = df['applications'].corr([df['acceptances'])
```

This resulted in a high positive correlation of 0.8017. This meant that as the number of applications increases, the number of acceptances increase as well. The scatterplot between these two is shown as the following:



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

In this question, I dropped the rows containing null values in 'applications', 'acceptances', and 'school_size'. I then set the application rate as:

appRate = df['applications']/df['school_size']

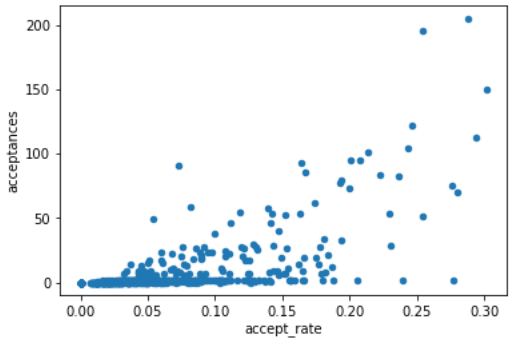
I then added this into the dataframe by running:

df.insert(0, 'accept_rate', acceptRate)

I found the correlation between acceptance and application rate by running:

correlation = df['acceptances'].corr(df['accept_rate])

This provided a correlation of 0.6587. This correlation is lower than the correlation between applications and acceptances. Since the correlation with the raw number of applications was higher, the raw number of applications would be a better predictor of admission to HSPHS.



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

In this question, I dropped the rows containing null values in ‘applications’, ‘acceptances’

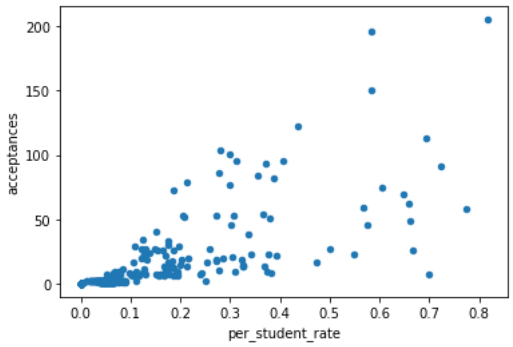
To find the per student odds, I ran the following command:

```
perStudentRate = df[‘acceptances’]/df[‘applications’]
```

Similarly to the previous question, I inserted this into the dataframe as ‘per_student_rate’

```
ind = perStudentRate.idxmax()
print(df['school_name'][ind])
print(perStudentRate.max())
```

The first command finds the index of the max perStudentRate. This index correlates to the schools with the highest per student rate. This school ended up being ‘THE CHRISTA MCAULIFFE SCHOOL\I.S. 187’. The last line prints out the perStudentRate as 0.8167.



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).

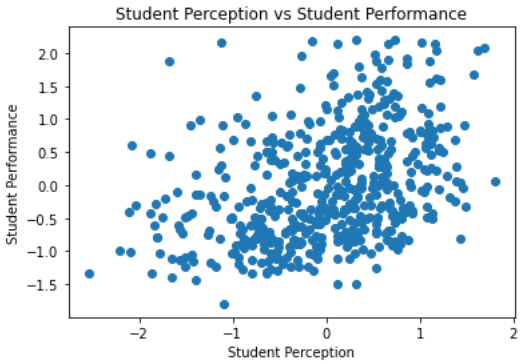
In this question, I dropped the rows containing null values based on the columns that we are using. I also normalized the data by z-scoring because the objective measures of achievements are not all scaled the same. For dimensional reduction, I ran the following code:

```
studentPerception = (df['rigorous_instruction'] + df['collaborative_teachers'] +
df['supportive_environment'] + df['effective_school_leadership'] +
df['strong_family_community_ties'] + df['trust'])/6

studentPerformance = (df['student_achievement'] + df['reading_scores_exceed'] +
df['math_scores_exceed'])/3
```

I then found the correlation between these two by running

```
correlation = studentPerception.corr(studentPerformance)
```

The correlation between how students perceive their school and how the school performs on objective measures of achievements is a moderate positive correlation of 0.3968.



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).

The hypothesis that I chose to test was high vs low poverty percentage based on the dependent variable of admissions to HSPHS. I divided high and low poverty percentage based on the median because this central tendency is least likely to be affected by outliers and is therefore more robust. I then ran the following code to run a Mann-Whitney U-test. I chose this test because we are comparing the two samples (high vs low poverty) to see if there is a significant difference between their admission rates to HSPHS. I also chose this test since the median is our measure of central tendency

```
for i in range(len(df['poverty_percent'])):
    if poverty.iloc[i] < medPov:
        lowPov.append(accepts.iloc[i])
    else:
```

```
        highPov.append(accepts.iloc[i])

stats.mannwhitneyu(lowPov, highPov)
```

We then get a p-value of 3.019831341650885e-22. Since this is lower than 0.05, we can reject the null hypothesis. This means that there is there is enough evidence to believe that there is a difference in high and low poverty rates in terms of acceptances.

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?

I had to z-score this data as well because the objective measures of achievement are not on the same scale.

```
achievements = (df['student_achievement'] + df['reading_scores_exceed'] +
df['math_scores_exceed'])/3

classAchievementCorr = df['avg_class_size'].corr(achievements)
classAcceptanceCorr = df['avg_class_size'].corr(df['acceptances'])

spendingAchievementCorr = df['per_pupil_spending'].corr(achievements)
spendingAcceptanceCorr = df['per_pupil_spending'].corr(df['acceptances'])
```

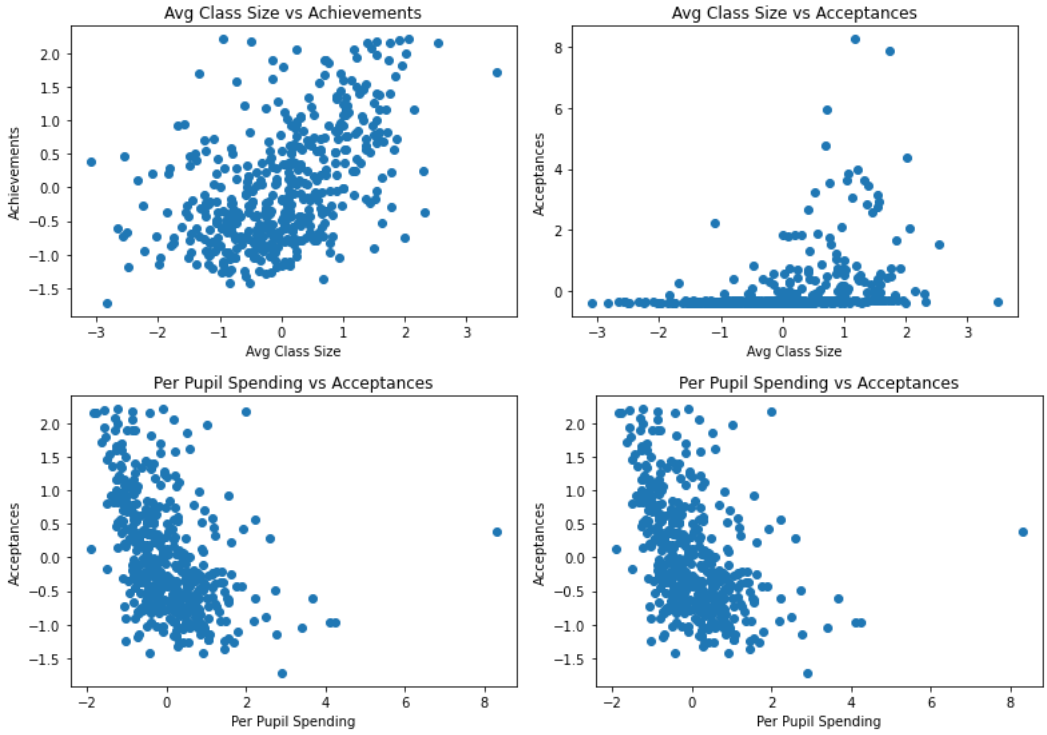
I found these four correlations. The correlation between them is as follows

Class Size vs Achievement: 0.5048063052192894
Class Size vs Acceptance: 0.348686320266411
Per Student Spending Vs Achievement: -0.44290689081093754
Per Student Spending vs Acceeptance: -0.3369726734526744

We can see that class size correlated with achievement and acceptances are both positive. This means that as the average class size increases, the objective measures of achievement also increase. Normally, people associate smaller class sizes being better, so there is no evidence that class size impacts objective measures of achievement.

For per student spending correlated with achievement and acceptances, they both are a negative correlation. As the per student spending increases, then both the objective measures of achievement and acceptances are decreasing. This would also be unexpected since people would

assume that having more resources available would help with achievement and acceptances. Therefore there is no evidence that per student spending impacts either of these two variables.

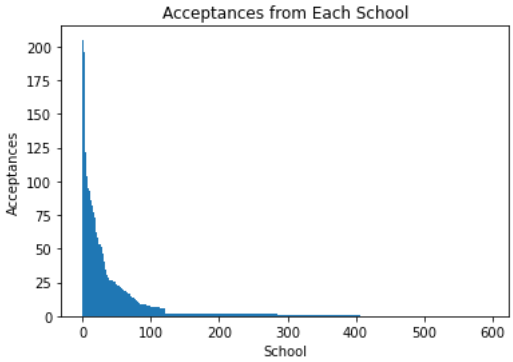


7) What proportion of schools accounts for 90% of all students accepted to HSPHS?

I dropped the rows containing null values based on the acceptances column.

```
accepts90 = df['acceptances'].sum()*0.9        #90% of the total acceptances

df = df.sort_values('acceptances', ascending=False)

items = df['acceptances'].iteritems()

soFar = 0
schoolCount = 0

for index,item in items:
    soFar += item
    schoolCount += 1
    if soFar >= accepts90:
        break;

print(schoolCount/len(df['acceptances']))
```

I First found 90% of all the students accepted to HSPHS by finding the sum of every row in the acceptances column and then multiplying it by 0.9. I then sorted these values from greatest to least (descending) because we are looking at the proportions of schools so we'll start from the schools with the most acceptances and move down the list. I then created an iterator to iterate through all the acceptances. With a for loop, I iterate through this iterator and add the running sum to a variable called soFar as well as increasing schoolCount by one each time. Once the running total is greater than or equal to the 90% of the total sum, it stops. Finding the proportion



by diving school count by the total number of acceptances, we get 0.2070. The following is the bar graph

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 dropped all the rows that contained a null value since I was running a regression that included all the columns.

```
#To find the independent variables in terms of acceptances
df1 = df.copy()
accepts = df1['acceptances']
df1 = df1.drop(["acceptances"], axis=1)

X = df1.loc[:, 'applications':'school_size']
y = accepts
```

```
X2 = sm.add_constant(X)
est = sm.OLS(y, X2)
est2 = est.fit()
print(est2.summary())
```

I decided to do a predictive model by running multiple regression to find what characteristics were most important. In terms of acceptances, I held “acceptances” as the dependent column and the rest of the column as the independent variables. When the summary of my multiple regression model was run, it achieved a R-squared of 0.704. This means that 70.4% of the variance can be explained by the models data. The independent variables that I found to be significant were “applications” and “school size” because they had a p-value less than 0.05. However, “poverty_percent” was also close since it came out with a p-value of 0.065 but it did not make the cut. I then ran another regression with these two variables only and our multiple regression equation would be as follows:

acceptances = 1.1076 + 0.3399(applications) + -0.0126(school size)

Similarly for objective measures of achievements, I ran a multiple regression with all the other columns to act as independent variables.

```
df2 = df.copy()
df2 = df2.drop(['school_name'], axis=1)
df2 = df2.drop(['dbn'], axis=1)
df2 = (df2 - df2.mean())/df2.std()

X_1 = df2.loc[:, 'applications':'school_size']
objAchievement = (df2['student_achievement'] + df2['reading_scores_exceed'] + df2['math_scores_exceed'])/3
y2 = objAchievement

#run model
X2 = sm.add_constant(X_1)
est_1 = sm.OLS(y2, X2)
est_2 = est_1.fit()

print(est_2.summary())
```

The significant variables that I found were “applications”, “rigorous instruction”, “supportive environment”, “disability percent”, and “poverty percent”. They had an R-squared value of 0.757 which meant 75.7% of the variance that can be explained by the model. The model is as follows:

Obj measure of achievement = -0.972 + 0.0024(applications) + 0.1504(rigorous_instruction) + 0.4066(supportive_environment) + -0.0212(disability_percent) + -0.193(poverty_percent)

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

We saw that there was a stronger positive correlation between the raw number of applications and acceptances than compared to correlation between application rate and acceptances. We know that the former had a stronger correlation because it was closer to 1. We then were able to see that the school with the best per student odds of getting accepted to HSPHS was "THE CHRISTA MCAULIFFE SCHOOL\I.S. 187". This school had 81.67% chance of getting accepted to HSPHS. By reducing the dimensions, we were able to find a positive moderate correlation of 0.3968 between how students perceive their school and how the school performs on objective measures of achievement. I then did a Mann Whitney U-test to test the hypothesis that I chose which was to see whether there was a difference in high vs low poverty percentage based on the dependent variable of admissions to HSPHS. As a result, we did see a p-value that was under 0.05 and was able to reject the null hypothesis. We were not able to find evidence that per student spending and average class sizes impacted the objective measures of achievements as per student spending had a negative correlation with both while average class sizes had a positive one. Next, we were able to see that 20.7% was the proportion of all schools that account for 90% of the total acceptance. Lastly, I did a multiple regression model to find which independent variables were significant in terms of sending students to HSPHS and objective measures of achievement. For the former, the variables that I found were significant were applications and school size as they had p-values under 0.05 and a R-squared of 0.704. For objective measures of achievement, the variables I found were applications, rigorous instruction, supportive environment, disability percent, and poverty percent. The R-squared value was 0.757. Overall, it seems that the characteristics that seem to be most relevant in determining acceptance would be applications and school size.

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) I would recommend that the school should encourage more students to apply to these schools to increase the number of applications. While doing so, they should also be looking at ways to improve objective measures of achievement.

b) To improve objective measures of achievement, I would recommend that schools increase their curriculum to make a bit more rigorous. While doing so, they should also increase the amount of support that students are able to receive by possibly holding more office hours.