Final Project

Net ID: zw2911

Bernice Wu

N-Number: N16499528

In this project, we explore a dataset provided by the New York City Department of Education. Using the following statistical analysis based on this dataset, we could have a general idea about what characteristics of NYC middle schools affect the admission rate to HSPHS and students' objective achievement most.

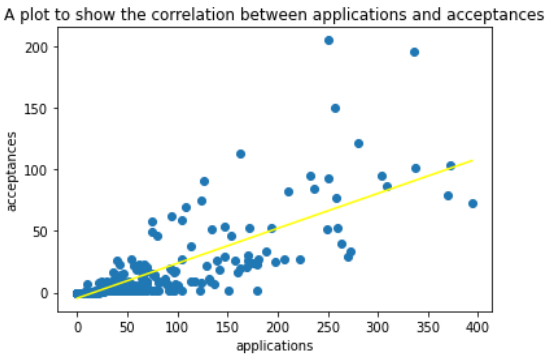
First and foremost, it is necessary to make a rough estimation of the dataset:

1. There are two groups of NYC middle school students, public and charter schools.
2. There are several missing data in this dataset.
3. The distributions of some data might be highly skewed.
4. Some characteristics may highly correlate with each other.

Accordingly, for the two different groups of schools, I make the statistical analysis separately. To deal with the missing data, I apply different methods for different questions. For questions 1)-5), I choose to drop the missing data since the missing data in the required dataset are in a small amount. For questions 6)-8), I import the IterativeImpute and fancyimpute class, which can model each feature with missing values as a function of other features, to impute the missing data and get a better inference. For the highly skewed data, I use z-scored and log-transforms to scale them. For the highly correlated characteristics, I use Principal Components Analysis (PCA) to achieve dimension reduction.

Q1: What is the correlation between the number of applications and admissions to HSPHS?

For this question, we could simply use numpy package to calculate the coefficient between the data in applications and admissions. And the correlation is 0.8.



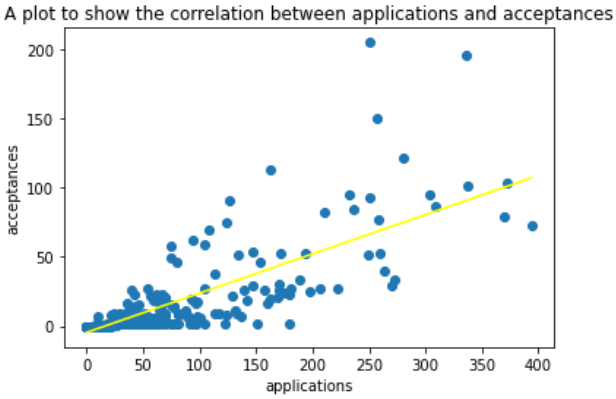

```
correlation = np.corrcoef(x,y)
print(correlation)
[[1.         0.80172654]
 [0.80172654 1.        ]]
```

Q2: What is a better predictor of admission to HSPHS? Raw number of applications or application *rate*?

To find the better predictor, we can fit the data with simple linear regression and analyze its coefficient of determination (r-squared).

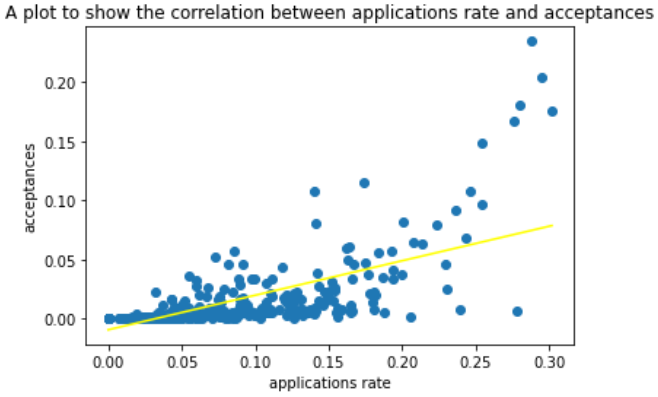
The COD for raw number of application is 0.643 while the COD for application rate is just 0.485, which means that using raw number of applications to predict the admission is closer to the real data. **Therefore, we can infer that the raw number of applications is the better predictor of admission to HSPHS.**

1) Raw number of applications:




```
==============================================================================
Dep. Variable:            acceptances   R-squared:                       0.643
Model:                            OLS   Adj. R-squared:                  0.642
Method:                 Least Squares   F-statistic:                     1062.
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          -4.5808      0.639     -7.172      0.000      -5.835      -3.326
applications    0.2840      0.009     32.586      0.000       0.267       0.301
```

2) Application rate:




```
==============================================================================
Dep. Variable:                      y   R-squared:                       0.485
Model:                            OLS   Adj. R-squared:                  0.484
Method:                 Least Squares   F-statistic:                     556.0
==============================================================================
```

```
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0092      0.001     -8.864      0.000      -0.011      -0.007
0              0.2918      0.012     23.580      0.000       0.268       0.316
```

Q3: Which school has the best *per student* odds of sending someone to HSPHS?

We first need to convert the number of acceptances to rate in order to obtain the *per student*'s performance, which also refers to the probability of admission; then we could calculate the odds.

We could sort the odds value in descending order to find the best school.

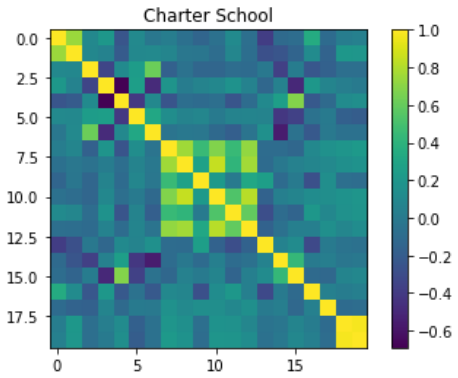
```
df3 = df[['school_name','applications','acceptances','school_size']].dropna()
rate = df3['acceptances'] / df3['school_size']
odds = rate / (1 - rate)
df3['odds'] = odds
df3 = df3.sort_values(by=['odds'], ascending=False)
```

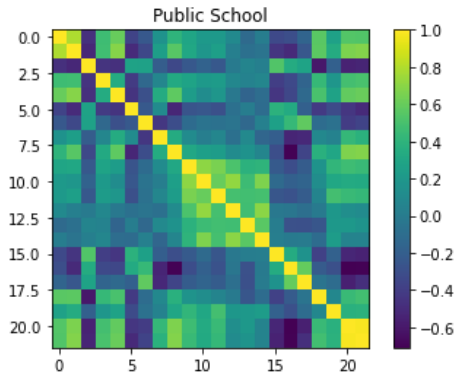
school_name	applications	acceptances	school_size	odds
THE CHRISTA MCAULIFFE SCHOOL\I.S. 187	251	205	873	0.306886
NEW YORK CITY LAB MIDDLE SCHOOL FOR COLLABORATIVE STUDIES	163	113	554	0.256236
M.S. 255 SALK SCHOOL OF SCIENCE	108	70	386	0.221519
J.H.S. 054 BOOKER T. WASHINGTON	257	150	852	0.213675
EAST SIDE MIDDLE SCHOOL	124	75	450	0.2

Therefore, THE CHRISTA MCAULIFFE SCHOOL has the best *per student* odds (~0.31) of sending someone to HSPHS.

Q4: 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).

As we learned in the lecture and recitation, we can firstly plot a correlation bar to see if there exist highly correlated variables. We can easily see a light square in the middle, which refers to the highly correlated school climate variables (columns L-Q); and another relatively light rectangle in the center of right side, which refers to the correlated objective achievement variables (columns V-X). Consequently, we could reduce the dimensions by doing Principal Components Analysis (PCA) so that we could grasp the general relationship between the school climate variables and students' objective achievements variables.

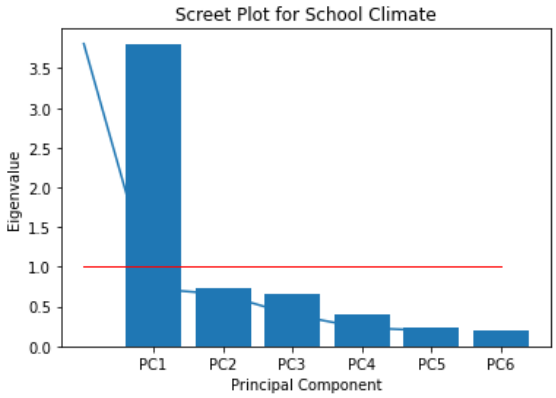




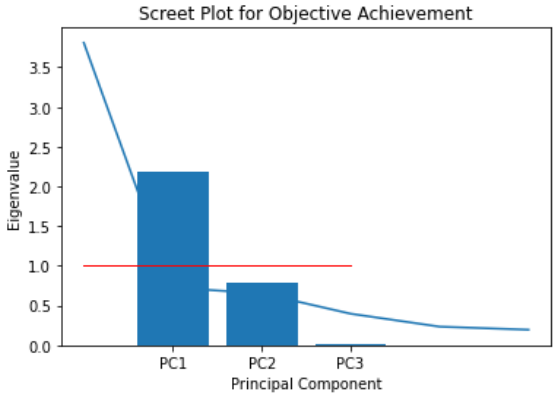
PCA on school climate:

PCA on objective achievements:

Based on the Kaiser criterion, we choose the first principal component dataset as a representative of the students' perception for the school climate.

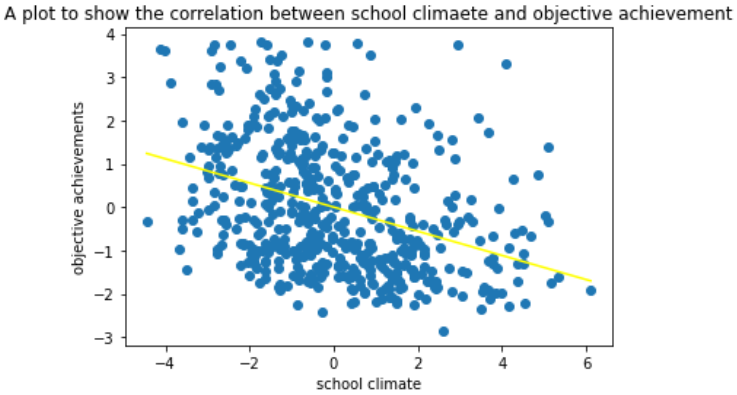


Similarly, we choose the first principal component to represent students' objective achievements.



We can then fit a simple linear regression on these two principal components and plot the data in order to see their correlation. Although the COD is relatively low, we can see directly from the plot that **the school climate and students' objective achievements are negatively correlated,** which means that the better the school climate is, the worse students' objective achievements will be.

```
==============================================================================
Dep. Variable:                      y   R-squared:                       0.135
Model:                            OLS   Adj. R-squared:                  0.133
Method:                 Least Squares   F-statistic:                     81.79
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       3.469e-18      0.060   5.77e-17      1.000      -0.118       0.118
x1            -0.2788      0.031     -9.044      0.000      -0.339      -0.218
```



Q5: 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 firstly transformed school_size variable into categorical one by evaluating them based on median number. Therefore, there are now two groups of school, large schools and small schools. I then did a t-test upon their admission rate to HSPHS to see if there is a significant difference between the means of two groups.

```
t,p = stats.ttest_ind(small_schools, large_schools) # independent t-test
print(p)
0.0013119187432157875
```

The p value is way smaller than 0.05, so we can reject the null hypothesis and draw a conclusion that the size of schools has an impact on its admission rate to HSPHS.

Q6: 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?

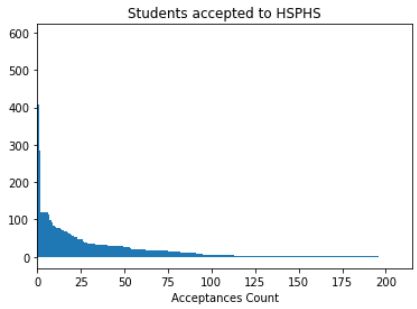
In the face of missing data in objective measures of achievement in charter schools, I decided to impute them based on the algorithms in IterativeImpute class imported from skit-learn package (learn it from https://scikit-learn.org/stable/). Similarly to Q5, I transformed the availability of material resources (per student spending) into categorical variable. Different from Q5, I classify them into 4 groups, which represent 4 levels of availability of material resources, so that I can do ANOVA or Kruskal-Wallis test upon their achievements to see if there is a significant difference between the means or medians of multiple groups.

```
f,p1 = stats.f_oneway(sample1,sample2,sample3,sample4)
print(p1)
6.901311074275004e-16
# kruskal-wallis on different groups
h,p2 = stats.kruskal(sample1,sample2,sample3,sample4) # 4 sample median
print(p2)
1.0356035842722143e-11
```

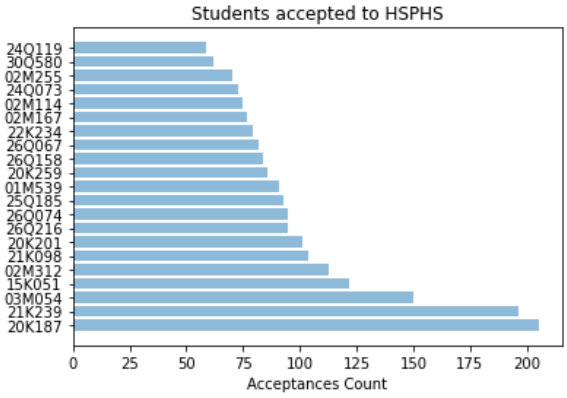
They are all significant enough to reject the null hypothesis. Accordingly, we can conclude that the availability of material resources has an influence upon the admission to HSPHS.

Q7: What proportion of schools accounts for 90% of all students accepted to HSPHS?

I firstly sort schools in a rank-ordered by decreasing number of acceptances of students to HSPHS. Then I calculated the number of 90% of all admitted students to HSPHS and ran for a loop to see the sum of acceptances from how many schools could match this number.



this bar graph shows the acceptances count in the whole dataset.



This bar graph only shows the top 20 schools' acceptances count.

```
df7 = df[['dbn','school_name','applications','acceptances']].dropna()
df7 = df7.sort_values(by=['acceptances'],ascending=False)
data = df7['acceptances'].values
num_acceptances = 0.9 * np.sum(data)
analysis = np.empty(len(data))
for i in range(len(data)):
    analysis[i] = np.sum(data[:i+1])
index = np.argmin(abs(analysis - num_acceptances))
proportion = index / len(data)
print(proportion)
0.2037037037037037
```

In conclusion, there are only 20% schools accounts for 90% of all students accepted to HSPHS. Just like a few percent of the population own the vast majority of the wealth.

Q8: 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 choose to make a prediction of the admission rate and the objective achievements of students based on the whole dataset.

Due to the complete absence of the data of availability of material resource, I firstly separate schools into two independent groups, public schools and charter schools, to achieve a better prediction.

For **Charter School**, we can simply drop the per_pupil_spending and avg_class_size data and impute the rest missing data by using k-NN classification from fancyimpute class. Due to the possible existence of highly-skewed data in dataset, we can log-transform data to rule out the effect of outliers. According to **Q4**, there are highly correlated variables which may contribute to the over-fitting issue when we make prediction, therefore, we need to replace those variables with principal components.

With all the preparation done, we could then use multiple linear regression to predict the admission rate and students' objective achievements.

Predict the Admission Rate: (compare them with the beta weight in the multiple regression)

```
Out[243]:
      Weight                   Name  Absolute Weight
0   1.459435           Applications         1.459435
7  -1.329456        poverty_percent         1.329456
9  -1.324633            school_size         1.324633
2  -0.462871          black_percent         0.462871
3  -0.270968       hispanic_percent         0.270968
11 -0.141939  Objective Achievement         0.141939
1  -0.072046          asian_percent         0.072046
5  -0.043345          white_percent         0.043345
6  -0.038010     disability_percent         0.038010
8  -0.023080            ESL_percent         0.023080
4   0.014383       multiple_percent         0.014383
10  0.013451         School Climate         0.013451
```

T**he most important school characteristics from charter school in terms of sending students to HSPHS are *Applications, poverty_percent* and *school_size*.**

Predict the Objective Achievements:

```
Out[244]:
Weight                      Name    Absolute Weight
8   1.583071     poverty_percent           1.583071
3  -0.979683       black_percent           0.979683
4  -0.885788    hispanic_percent           0.885788
0   0.587942        applications           0.587942
1  -0.518654         acceptances           0.518654
10 -0.439435         school_size           0.439435
5   0.242880    multiple_percent           0.242880
7  -0.165001  disability_percent           0.165001
11  0.095575      School Climate           0.095575
9  -0.072723         ESL_percent           0.072723
2  -0.030577       asian_percent           0.030577
6  -0.021669       white_percent           0.021669
```

T**he most important school characteristics from charter school in terms of achieving high scores on objective measures of achievement are *poverty_percent*, *black_percent* and *hispanic_percent*.**

For **Public Schools**, the statistical analysis is identical as for charter schools. The only exception is to take the availability of material resource into account.

Predict the Admission Rate: (only take the top 3 beta weight)

```
Out[257]:
Weight                         Name    Absolute Weight
9  -1.393537        poverty_percent           1.393537
0   0.960591           applications           0.960591
1  -0.321189     per_pupil_spending           0.321189
```

T**he most important school characteristics from public school in terms of sending students to HSPHS are *poverty_percent*, *applications* and *per_pupil_spending*.**

Predict the Objective Achievements: (only take the top 3 beta weight)

```
Out[259]:
Weight                       Name    Absolute Weight
8  -1.190197      poverty_percent           1.190197
7  -0.497311   disability_percent           0.497311
9  -0.312469          ESL_percent           0.312469
```

T**he most important school characteristics from public school in terms of achieving high scores on objective measures of achievement are *poverty_percent*, *disability_percent* and *ESL_percent*.**

Q9: what school characteristics seem to be most relevant in determining acceptance of their students to HSPHS?

To public schools, I think the most relevant school characteristics in determining acceptance of their students to HSPHS are poverty percent, applications and per_pupil_spending, which indicates that economic conditions for education and life as well as applications' rate have the most important impact on their admissions' rate.

To charter schools, applications, poverty_percent and school_size account for the most relevant school characteristics in terms of the acceptances to HSPHS, uncovering that their admissions' rate are mainly affected by applications' rate, basic economic conditions and the school size.

Q10: 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.

Generally speaking, I think the Department of Education should increase educational expenditures and simultaneously reduce schools' tuition and fees, which can ease the economic burden on ordinary students and thus enabling them to focus more on the study. I also suggest that the department encourage students to apply for HSPHS and be confident of their abilities. The schools should also embrace ethnic diversity and care for the disabled groups, thus creating a more harmonious environment for students to enjoy the study life. Accordingly, the admission rate to HSPHS and students' objective achievements can then be significantly improved.