Introduction by Shimun Naher

In this final project, I will be exploring a dataset provided by the New York City Department of Education. In my dataset, there was some dataset missing and so I went in the direction of removing the missing values. I removed the missing values for columns containing large numbers of them because I felt it was easier to analyze the data without them. For question 4, I used PCA to handle the dimension reduction of the data.

In [8]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df=pd.read_csv("middleSchoolData.csv")

df.isnull().sum()

df['per_pupil_spending'] = df.dropna()['per_pupil_spending']
df['avg_class_size'] = df.dropna()['avg_class_size']
df['school_size'] = df.dropna()['school_size']
```

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

In [9]:

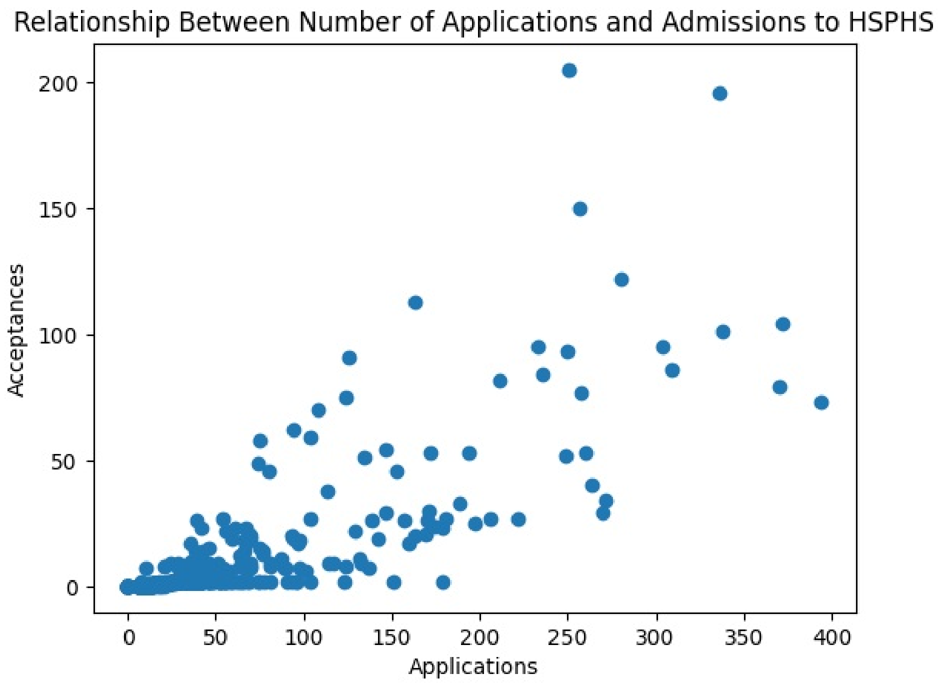
```
df['acceptances'].corr(df['applications'])
```

Out[9]:

```
0.8017265370719315
```

In [10]:

```
x = df['applications']
y = df['acceptances']

plt.scatter(x,y)
plt.xlabel("Applications")
plt.ylabel("Acceptances")
plt.title("Relationship Between Number of Applications and Admissions to HSPHS")
plt.show()
```



To find the correlation between the number of applications and admissions to HSPHS, I found the correlation coefficient between the columns "applications" and "acceptances". The value I found for the correlation is approximantely 0.80. I also created a scatter plot displaying the data points for the number of applications (on the x-axis) and the number of acceptances (on the y-axis). From the correlation coefficient value and the scatter plot it's clear that there is a strong correlation between the two variables and that the number of acceptances is increasing when the number of applications is increasing. Additionally, the correlation between the two variables is positive because the data points go from being mostly clustered together to the left of the graph to increasing more on the right of the graph.

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

In [11]:
```
df['application_rate']=df['applications']/df['school_size']

df['acceptances'].corr(df['application_rate'])
```

Out[11]:
```
0.6865549148749962
```

The application rate was not provided in the data so to find the values for it I divided the number of applications by size of the school. To determine whether raw number of applications or application rate is a better predictor of admission to HSPHS I calculated the correlation coefficient value between the number of acceptances and the application rate. I did not calculate the correlation between the number of applications and acceptances because that value is already calculated in Q1. The correlation value found between application rate and acceptances is approximately 0.69 which is smaller than 0.80 from the previous problem. Though, there may be other factors not accounted for, because the value for raw number of applications correlated with acceptances to HSPHS is greater, I conclude that the raw number of applications is a better predictor of admission to HSPHS than application rate.

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

In [12]:
```
df['per_student_odds']= (df['acceptances']/df['applications'])

df[['school_name','applications','acceptances','per_student_odds']].sort_values('per_student_odds',ascending=Fals
```

Out[12]:

	school_name	applications	acceptances	per_student_odds
304	THE CHRISTA MCAULIFFE SCHOOL\I.S. 187	251	205	0.816733
47	THE ANDERSON SCHOOL	75	58	0.773333
8	NEW EXPLORATIONS INTO SCIENCE, TECHNOLOGY AND ...	126	91	0.722222
50	SPECIAL MUSIC SCHOOL	10	7	0.700000
22	NEW YORK CITY LAB MIDDLE SCHOOL FOR COLLABORAT...	163	113	0.693252
...	...	...	...	...
446	P.S. 111 JACOB BLACKWELL	0	0	NaN
531	CAPITAL PREPARATORY (CP) HARLEM CHARTER SCHOOL	0	0	NaN
537	NEW YORK CENTER FOR AUTISM CHARTER SCHOOL	0	0	NaN
541	NEW HEIGHTS ACADEMY CHARTER SCHOOL	0	0	NaN
568	BRONX LIGHTHOUSE CHARTER SCHOOL	0	0	NaN

594 rows × 4 columns

The way that I interpreted the phrase "per student odds" for this problem is the number of students who were admitted to HSPHS out of all of the students who applied for admission from each school. This question could also be interpreted as the number of acceptances out of the population of each school but I was not sure if this is what the question was seeking so I went with the former. To calculate my per student odds value I created a variable with the same name and set that equal to the number of acceptances divided by the number of applications. With that value for each school, I outputted a table with the columns that identified the name of the school, the number of applications, the number of acceptances, and the per student odds. I then ranked the per student odds values from highest to lowest to find the school with the highest per student odds. The school that has the best per student odds of sending someone to HSPHS is The Christa Mcauliffe School/I.S. 187.

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).

In [13]:
```
from sklearn.decomposition import PCA

perception = df.dropna()[['rigorous_instruction','collaborative_teachers','supportive_environment',
    'effective_school_leadership','strong_family_community_ties','trust'
```

```
    ]]
objective = df.dropna()[['student_achievement','reading_scores_exceed','math_scores_exceed']]

pc = PCA(1).fit(perception)

perception_reduced = pc.transform(perception)

print(pc.explained_variance_ratio_)

cp = PCA(1).fit(objective)

objective_reduced = cp.transform(objective)

print(cp.explained_variance_ratio_)

np.corrcoef(perception_reduced.reshape(1,-1),objective_reduced.reshape(1,-1))
```

```
[0.63550425]
[0.87851003]
```

Out[13]:

```
array([[1.        , 0.4148258],
       [0.4148258, 1.        ]])
```

For this problem, I first used Principal Component Analysis to reduce the dimensionality of the data sets because this question involved multiple columns compared to the previous questions. Columns L-Q are represented by perception and columns V-X are represented by objective. The first two values printed represent how much of the data is represented after application PCA to the columns. Once I completed the dimension reduction aspect of the problem I reshaped perception and reduction data to find the correlation coefficient. The correlation coefficient value is approximately 0.41 which tells me that there is a weak, positive relationship between how students perceive their school and how the school performs on objective measures of achievement.

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).

In [14]:

```
df['category']=np.where((df['school_size'])>=df['school_size'].median(),'big','small')
df.groupby(['category']).count()
```

Out[14]:

	dbn	school_name	applications	acceptances	per_pupil_spending	avg_class_size	asian_percent	black_percent	hispanic_percent	mu
category										
big	225	225	225	225	225	225	225	225	225	
small	369	369	369	369	224	224	367	367	367	

2 rows × 26 columns

My hypothesis for Q5 is that smaller schools will perform better than larger schools in having more students admitted to HSPHS. To approach this problem, I created a column called 'category' and split all of the middle schools into the groups 'big' or 'small' based on the median of the school size. Once I created the two groups I produced a table with a count of how many students were accepted to HSPHS in addition to the other variables provided by the data. From observation, it's clear that the numbers between big and small schools are not very close to one another and so the size of a school does show to change the performance of students on getting admitted to HSPHS. In this case, smaller schools tend to have more admissions to HSPHS.

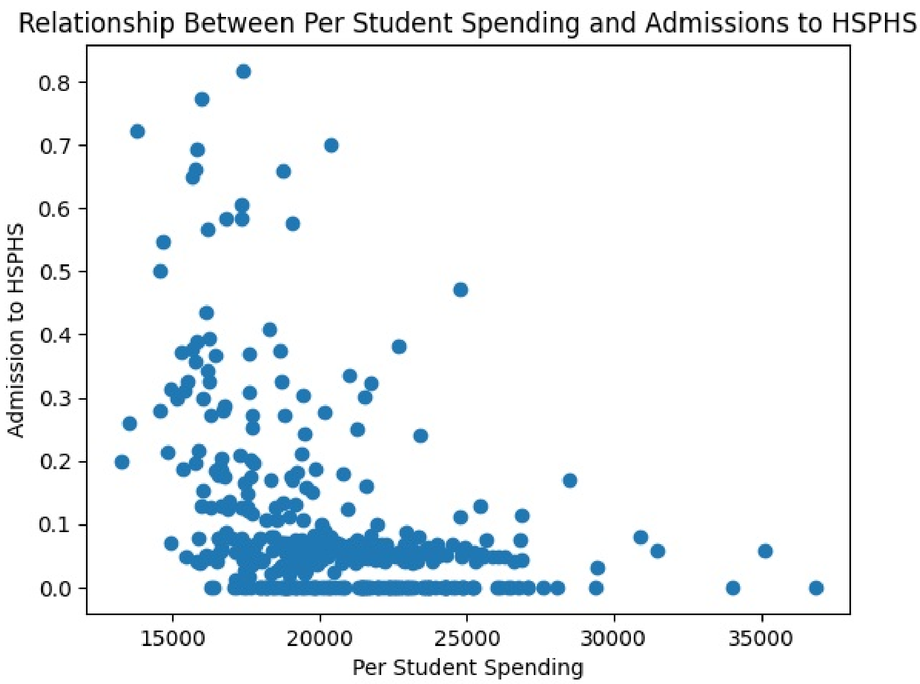
Q6: Is there any evidence that the availability of material resources (e.g. per student spending or class size) impacts objective measures of achievement or admission to HSPHS?

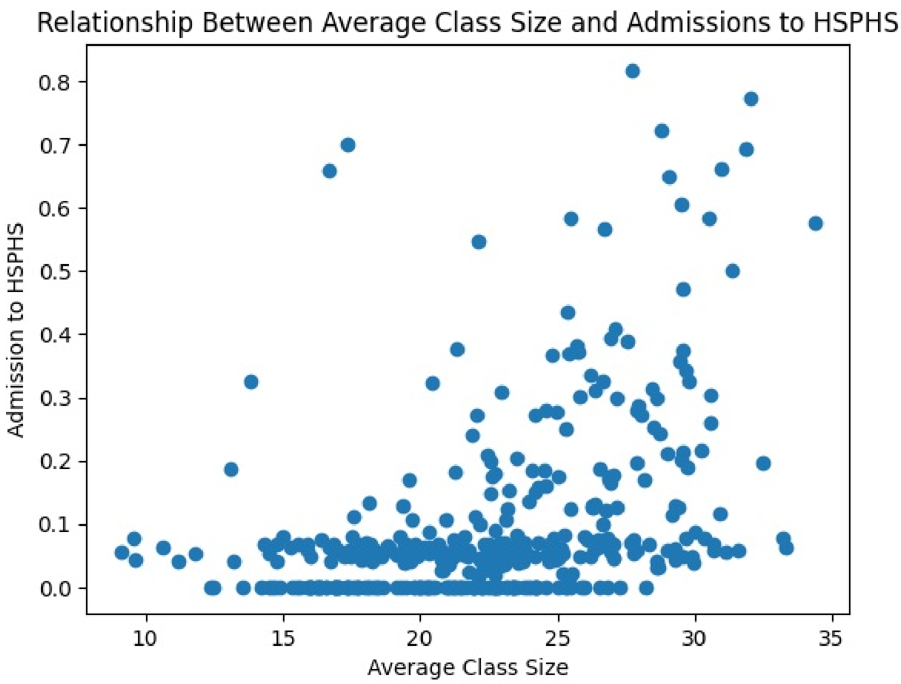
In [15]:

```
ans1 = df[['per_pupil_spending']].corrwith(df['acceptances']/df['applications'])
print(ans1)

clas1 = df[['avg_class_size']].corrwith(df['acceptances']/df['applications'])
```

```
print(clas1)

ansx = df['per_pupil_spending']
ansy = df['acceptances']/df['applications']

plt.scatter(ansx,ansy)
plt.xlabel("Per Student Spending")
plt.ylabel("Admission to HSPHS")
plt.title("Relationship Between Per Student Spending and Admissions to HSPHS")
plt.show()


clasx = df['avg_class_size']
clasy = df['acceptances']/df['applications']

plt.scatter(clasx,clasy)
plt.xlabel("Average Class Size")
plt.ylabel("Admission to HSPHS")
plt.title("Relationship Between Average Class Size and Admissions to HSPHS")
plt.show()
```

```
per_pupil_spending   -0.394344
dtype: float64
avg_class_size    0.418502
dtype: float64
```





For this problem, I searched for evidence for per student spending impacting admission to HSPHS in addition to class size impacting admission to HSPHS. The variable ans1 represents the correlation value between per student spending and the number of students accepted to HSPHS divided by the number of applications. The variable clas1 represents the correlation value between average class size and the number of students accepted to HSPHS divided by the number of applications. For ans1, my correlation value is approximately -0.39 which tells me that there is a weak, negative relationship between per student spending and admission to HSPHS, also shown by the corresponding graph. From this, I know that there is no evidence that the per student spending impacts admission to HSPHS. For clas1, my correlation value is approximately 0.42 which tells me that there is a moderate, positive relationship between average class size and admission to HSPHS, also shown by the corresponding graph. From this, I know that there is little evidence that the size of a class impacts the number of students admitted to HSPHS.

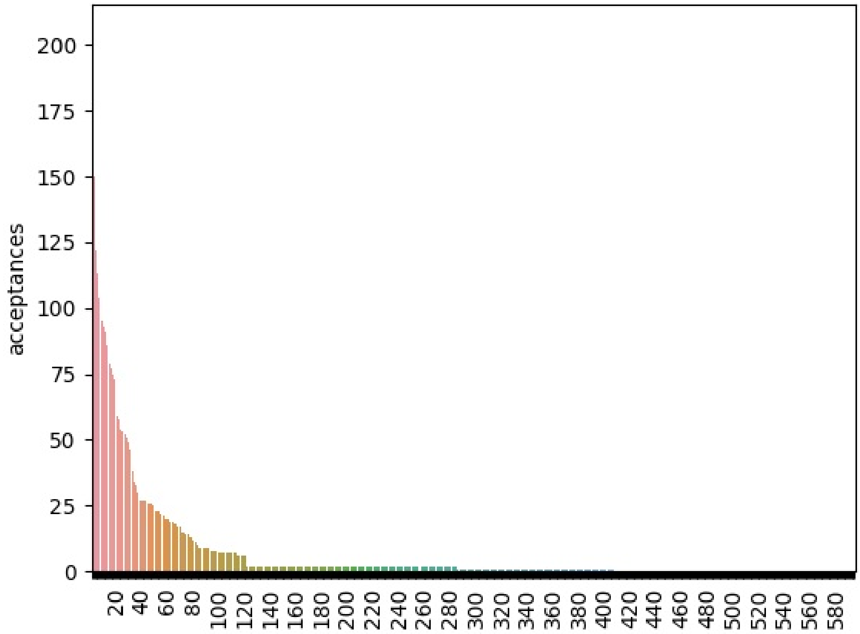
Q7: What proportion of schools accounts for 90% of all students accepted to HSPHS?

In [16]:

```
import seaborn as sns

df['Total']=df['acceptances'].sum()

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

df['cumsum']=df['acceptances'].cumsum()

df['Percentage']=(df['cumsum']/df['Total'])*100

e=df[df["Percentage"]<=90]

print(e)

data = df.sort_values('acceptances',ascending=False).reset_index()
data['num']= data.index + 1

data.loc[data['num']%20!=0,'num'] = " "
data.loc[data['num']==" ",'num']= data.loc[data['num']==" ",'num']* data.loc[data['num']==" ",'num'].index
data.loc[data['num']==" ",'num']= data.loc[data['num']==" ",'num']* data.loc[data['num']==" ",'num'].index
sns.barplot(data=data,x='num',y='acceptances')
plt.xticks(rotation=90)

plt.show()
```

```
        dbn                                       school_name  applications  \
304  20K187             THE CHRISTA MCAULIFFE SCHOOL\I.S. 187           251
324  21K239         MARK TWAIN I.S. 239 FOR THE GIFTED & TALENTED           336
33   03M054                    J.H.S. 054 BOOKER T. WASHINGTON           257
241  15K051                       M.S. 51 WILLIAM ALEXANDER           280
22   02M312  NEW YORK CITY LAB MIDDLE SCHOOL FOR COLLABORAT...           163
..      ...                                               ...           ...
45   03M291                          WEST END SECONDARY SCHOOL            31
468  31R051                           I.S. 051 EDWIN MARKHAM           101
587  84X494             SUCCESS ACADEMY CHARTER SCHOOL - BRONX 2            36
186  11X144                        J.H.S. 144 MICHELANGELO            42
407  27Q282          KNOWLEDGE AND POWER PREPARATORY ACADEMY VI            42

     acceptances  per_pupil_spending  avg_class_size  asian_percent  \
304          205             17403.0           27.71           67.5
324          196             16814.0           30.51           27.8
33           150             17359.0           25.47           11.0
241          122             16145.0           25.36           16.4
22           113             15853.0           31.83           55.1
..           ...                 ...             ...            ...
45             6                 NaN             NaN            8.0
468            6             19320.0           19.47            5.6
587            6                 NaN             NaN            1.3
186            2             21736.0           20.97            2.8
407            2             16253.0           26.20            3.8

     black_percent  hispanic_percent  multiple_percent  ...  school_size  \
304            1.3               6.8               0.6  ...        873.0
324            6.9               5.7               8.2  ...       1322.0
33             8.6              14.7               8.0  ...        852.0
241            7.6              20.4               4.7  ...       1136.0
22             2.0               6.5               6.5  ...        554.0
..             ...               ...               ...  ...          ...
45             5.9              18.9               4.4  ...          NaN
468           23.6              44.3               1.3  ...       1315.0
587           61.3              35.9               0.5  ...          NaN
186           60.2              32.5               0.4  ...        492.0
407           50.0              42.8               1.1  ...        264.0

     student_achievement  reading_scores_exceed  math_scores_exceed  \
304                 4.36                   0.90                0.90
324                 4.16                   0.88                0.88
33                  4.14                   0.89                0.88
241                 4.05                   0.83                0.83
```

```
22                 4.34                 0.88                0.89
..                  ...                  ...                 ...
45                 4.19                 0.80                0.79
468                3.29                 0.66                0.71
587                 NaN                 0.79                0.80
186                3.86                 0.29                0.19
407                4.64                 0.44                0.34

     application_rate  per_student_odds  category  Total  cumsum  Percentage
304          0.287514          0.816733       big   4461     205    4.595382
324          0.254160          0.583333       big   4461     401    8.989016
33           0.301643          0.583658       big   4461     551   12.351491
241          0.246479          0.435714       big   4461     673   15.086304
22           0.294224          0.693252       big   4461     786   17.619368
..                ...               ...       ...    ...     ...         ...
45                NaN          0.193548     small   4461    3998   89.621161
468          0.076806          0.059406       big   4461    4004   89.755660
587               NaN          0.166667     small   4461    4010   89.890159
186          0.085366          0.047619     small   4461    4012   89.934992
407          0.159091          0.047619     small   4461    4014   89.979825

[122 rows x 30 columns]
```



num

For this question, I produced a table with a column containing the percentage of students admitted to HSPHS. After doing that, I limited the number of rows produced by the table to only account for 90% of all students accepted. The number of rows produced is 122 listed below the table which represents the number of schools that account for 90% of students accepted. The total number of schools represented in this data is 594 so the proportion of schools that represents 90% of all students accepted is 122/594 or approximately 20.54%. Below the table, I also created a bar graph of schools, rank-ordered by decreasing number of acceptances of students to HSPHS. The numbers on the x-axis represent the 'dbn' of the schools.

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?

In [17]:

```
import statsmodels.api as sm
#part A
x = df.dropna().drop(['acceptances','per_student_odds','school_name','dbn','category'],axis=1)

y = df.dropna()['per_student_odds']

x = sm.add_constant(x)

results = sm.OLS(y,x).fit()

print(results.summary())

#part B

x1 = df.dropna().drop(['acceptances','per_student_odds','school_name','dbn',
                       'student_achievement','reading_scores_exceed','math_scores_exceed','category'],axis=1)

x1 = sm.add_constant(x1)

y1 = objective_reduced

results1 = sm.OLS(y1,x1).fit()

print(results1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       per_student_odds   R-squared:                       0.882
Model:                            OLS   Adj. R-squared:                  0.875
Method:                 Least Squares   F-statistic:                     122.0
```

```
Date:                Thu, 19 Aug 2021   Prob (F-statistic):          1.01e-158
Time:                        19:31:45   Log-Likelihood:                 636.32
No. Observations:                 400   AIC:                            -1225.
Df Residuals:                     376   BIC:                            -1129.
Df Model:                          23
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
applications                    -0.0020      0.000    -14.529      0.000      -0.002      -0.002
per_pupil_spending            5.304e-07   1.12e-06      0.474      0.635   -1.67e-06    2.73e-06
avg_class_size                  -0.0010      0.001     -1.354      0.177      -0.002       0.000
asian_percent                    0.0513      0.039      1.314      0.190      -0.025       0.128
black_percent                    0.0511      0.039      1.310      0.191      -0.026       0.128
hispanic_percent                 0.0514      0.039      1.317      0.189      -0.025       0.128
multiple_percent                 0.0565      0.039      1.450      0.148      -0.020       0.133
white_percent                    0.0511      0.039      1.311      0.191      -0.026       0.128
rigorous_instruction             0.0155      0.006      2.416      0.016       0.003       0.028
collaborative_teachers          -0.0185      0.008     -2.468      0.014      -0.033      -0.004
supportive_environment          -0.0023      0.007     -0.310      0.757      -0.017       0.012
effective_school_leadership     -0.0131      0.008     -1.703      0.089      -0.028       0.002
strong_family_community_ties    -0.0062      0.006     -1.028      0.304      -0.018       0.006
trust                            0.0195      0.008      2.369      0.018       0.003       0.036
disability_percent              -0.0002      0.001     -0.318      0.751      -0.001       0.001
poverty_percent                 -0.0006      0.000     -1.777      0.076      -0.001    6.08e-05
ESL_percent                   6.912e-05      0.000      0.188      0.851      -0.001       0.001
school_size                   5.745e-05   1.54e-05      3.720      0.000    2.71e-05    8.78e-05
student_achievement              0.0123      0.005      2.546      0.011       0.003       0.022
reading_scores_exceed           -0.0327      0.100     -0.327      0.744      -0.229       0.164
math_scores_exceed               0.0699      0.086      0.811      0.418      -0.100       0.239
application_rate                 0.4141      0.105      3.962      0.000       0.209       0.620
Total                           -0.0009      0.001     -1.032      0.303      -0.003       0.001
cumsum                          -0.0002     8.5e-06    -28.184      0.000      -0.000      -0.000
Percentage                   -5.367e-06     1.9e-07    -28.184      0.000   -5.74e-06   -4.99e-06
==============================================================================
Omnibus:                      186.129   Durbin-Watson:                   1.326
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1985.840
Skew:                           1.689   Prob(JB):                         0.00
Kurtosis:                      13.380   Cond. No.                     9.68e+17
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1.97e-25. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.062
Model:                            OLS   Adj. R-squared:                  0.012
Method:                 Least Squares   F-statistic:                     1.244
Date:                Thu, 19 Aug 2021   Prob (F-statistic):              0.214
Time:                        19:31:45   Log-Likelihood:                -405.64
No. Observations:                 400   AIC:                             853.3
Df Residuals:                     379   BIC:                             937.1
Df Model:                          20
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
applications                    -0.0048      0.002     -2.609      0.009      -0.008      -0.001
per_pupil_spending           -7.302e-06   1.48e-05     -0.492      0.623   -3.65e-05    2.19e-05
avg_class_size                  -0.0042      0.010     -0.428      0.669      -0.023       0.015
asian_percent                    0.4386      0.523      0.838      0.402      -0.590       1.468
black_percent                    0.4412      0.523      0.844      0.399      -0.587       1.470
hispanic_percent                 0.4396      0.523      0.840      0.401      -0.589       1.468
multiple_percent                 0.4551      0.523      0.871      0.385      -0.573       1.483
white_percent                    0.4415      0.523      0.844      0.399      -0.587       1.470
rigorous_instruction            -0.0700      0.085     -0.823      0.411      -0.237       0.097
collaborative_teachers           0.0682      0.101      0.678      0.498      -0.130       0.266
supportive_environment           0.0485      0.094      0.516      0.606      -0.136       0.234
effective_school_leadership     -0.0229      0.103     -0.222      0.824      -0.226       0.180
strong_family_community_ties    -0.0044      0.080     -0.055      0.956      -0.163       0.154
trust                           -0.0088      0.108     -0.081      0.935      -0.222       0.204
disability_percent               0.0030      0.008      0.402      0.688      -0.012       0.018
poverty_percent                  0.0034      0.004      0.806      0.421      -0.005       0.012
ESL_percent                     -0.0024      0.004     -0.534      0.593      -0.011       0.006
school_size                      0.0004      0.000      1.743      0.082   -4.64e-05       0.001
application_rate                 0.8120      1.389      0.585      0.559      -1.919       3.542
Total                           -0.0098      0.012     -0.834      0.405      -0.033       0.013
cumsum                          -0.0001      0.000     -1.213      0.226      -0.000    8.48e-05
Percentage                   -3.062e-06   2.52e-06     -1.213      0.226   -8.02e-06     1.9e-06
==============================================================================
Omnibus:                        6.258   Durbin-Watson:                   1.783
Prob(Omnibus):                  0.044   Jarque-Bera (JB):                4.295
Skew:                           0.099   Prob(JB):                        0.117
Kurtosis:                       2.533   Cond. No.                     7.94e+17
==============================================================================

Notes:
```

```
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 2.93e-25. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

```
/Users/rudra/opt/anaconda3/lib/python3.7/site-packages/statsmodels/tsa/tsatools.py:142: FutureWarning: In a futur
e version of pandas all arguments of concat except for the argument 'objs' will be keyword-only
  x = pd.concat(x[::order], 1)
/Users/rudra/opt/anaconda3/lib/python3.7/site-packages/statsmodels/tsa/tsatools.py:142: FutureWarning: In a futur
e version of pandas all arguments of concat except for the argument 'objs' will be keyword-only
  x = pd.concat(x[::order], 1)
```

To answer Q8 I used two separate statsmodels to receive the ordinary least squares regression outputs for Part A and Part B. To determine what school characteristics are most important in sending students to HSPHS I looked for columns that had a p-value less than 0.05 in Part A. From my observations, I found that number of applications, rigorous instruction, collaborative teachers, trust, school size, student achievement, and application rate are school characteristics that are important for sending students to HSPHS. To determine what school characteristics are important in achieving high scores on objective measures of achievement I looked for columns that had a p-value less than 0.05 in Part B. From my observations, I found that number of applications to HSPHS are important for achieving high scores on objective measures of achievement.

Q9: Write an overall summary of your findings – what school characteristics seem to be most relevant in determining acceptance of their students to HSPHS?

To determine what school characteristics seem to be most relevant in determining acceptance of their students to HSPHS I based my answer on my responses for questions 5, 6, and 8. From question 5, it was shown that smaller schools are more likely to have their students admitted to HSPHS. Potential reasons for this could be that there is more undivided attention on each students' performance and students can focus on learning relevant topics for the exam better. From question 6, it was shown that there is a relationship between the average class size and the performance of students on an exam. The amount of money spent on each student was shown to be irrelevant. From question 8, it was shown that there is a relationship between the number of applications, rigorous instruction, collaborative teachers, trust, school size, student achievement, and application rate for sending students to HSPHS. Based on these findings, I believe that the school characteristics most relevant in determining acceptance of their students to HSPHS are the number of students who apply, the size of the school, the size of their classes, the rigor of instruction, the collaborative teachers, student achievement, and the trust relationship between students and their school.

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.

Part A: To answer this part of the question my answer will heavily rely on my answer to Q9 because the factors that are determined to send more students to HSPHS is listed there. To improve the number of students sent to HSPHS, I would recommend the New York City Department of Education to improve the quality of teachers at their schools to ensure that all of the students are provided with an equal amount of good education that prepares them for the exam. I would also recommend the rigor of classes to increase so that students learn more and are tested on the things they learned to make sure they actually understand it. Additionally, I would recommend telling more students about HSPHS and pushing them to apply because applying and knowing if they will get admitted is better than not even trying. If possible, I would also recommend having larger classes focused on rigorous courses but a smaller school so that each classroom of students receives the attention and quality education they deserve. Lastly, I would recommend reassessing the environment of the school and seeing if it fosters a learning environment for students by allowing them to openly ask questions to obtain trust in the people teaching them.

Part B: Based on the results from this study, the factors I would recommend to improve measures are similar to Part A. I would recommend increasing the rigor of the coursework because students will be challenged to learn more topics and be better prepared for exams in the relevant subjects. I would also recommend more collaboration among teachers because they are the primary sources of education for the students. Adding to that, I would recommend schools focus more on the environment students are in to create one that is trustworthy and supportive to foster engagement and enthusiasm among students.