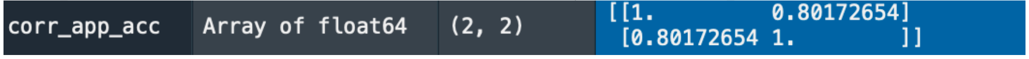
Report of Middle School Data in New York

--- Moon Liang (xl3265)

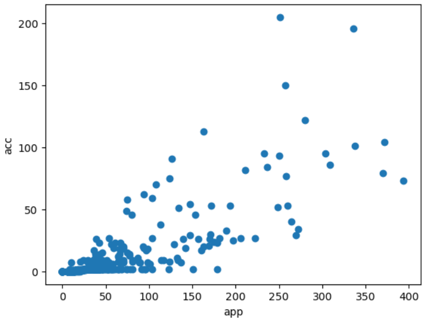
The PCA method is used to do the dimension reduction for the data that is use in this Report, specifically on factors of school climates (L-Q) and objective student achievement measures (V-X). Considering there are missing data, nans are first imputed to take place of the missing data to make sure that statistics are computed with as much data as possible. With problems of which data cannot be computed with NaNs, data is processed by cleaning the rows that contain any NaNs. As Charter Schools have data that is missing systematically, they have been considered as another category and discussed separately in specific questions. Below are the specifics of this report.

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

According to the data, the correlation between the number of applications and acceptances to HSPHS is 0.8017 as shown by the scatterplot below:

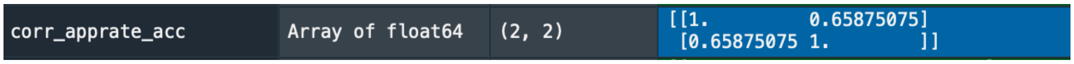


```
corr_app_acc    Array of float64    (2, 2)    [[1.          0.80172654]
                                               [0.80172654 1.        ]]
```

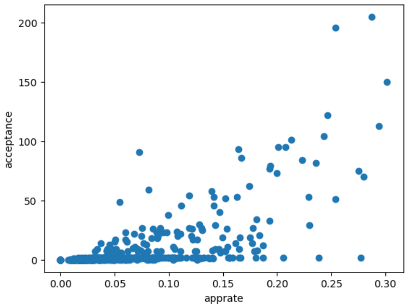


2. **What is a better predictor of admission to HSPHS? Raw number of applications or application *rate*?**

The correlation between the application rate (#application/schoolsize) and number of acceptances to HSPHS is 0.66 which is less than the correlation between the raw number of applications and the number of acceptances to HSPHS (0.8). Therefore, raw number of applications shows a strong relationship with number of acceptances to HSPHS, which indicates that it is a better predictor of admission.

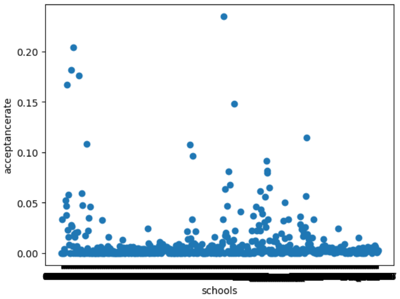


```
corr_apprate_acc    Array of float64    (2, 2)    [[1.          0.65875075]
                                                   [0.65875075 1.        ]]
```



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

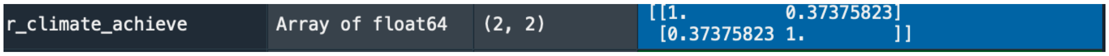
According to the data, by getting the acceptance rate(#acceptances/schoolsize) of each school, the highest acceptance rate among all schools is 0.235=205/873, which is data comes from The Christa Mcaulife School\I.S. 187 as it is positioned #305 in the datafile. Therefore, The Christa Mcaulife School has the best per student odds of sending someone to HSPHS as 205 out of 873 its total student body got accepted.



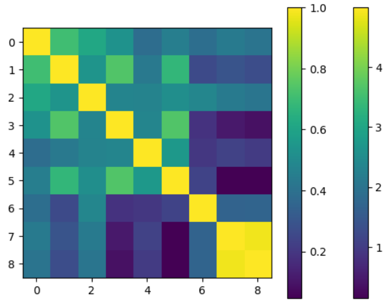
```
best    float    1    0.23482245131729668
```

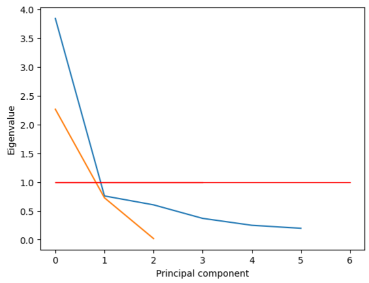
4. **Is there a relationship between how students perceive their school (as reported in columns. L-Q) and how the school performs on objective measures of achievement (as noted in columns V-X).**

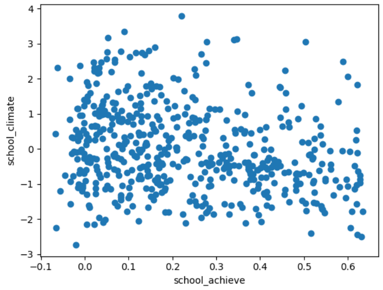
From the PCA process, the correlation matrix and the scree plot has shown that there is a one factor in either group that explains most of the variance as there is one factor above the eigenvalue of 1 for both groups. From dimension reduction, then a factor of school climate is generated that account for factors in columns, L-Q, and another factor of objective student achievement is generated accounts for factors in columns, V-X. By calculating the correlation coefficient, the value of 0.37 shows that these two factors are not as strongly correlated and there is barely a relationship between how students perceive their school and how the school performs, which can also be seen from the scatterplot below.



```
r_climate_achieve    Array of float64    (2, 2)    [[1.         0.37375823]
                                                    [0.37375823 1.        ]]
```







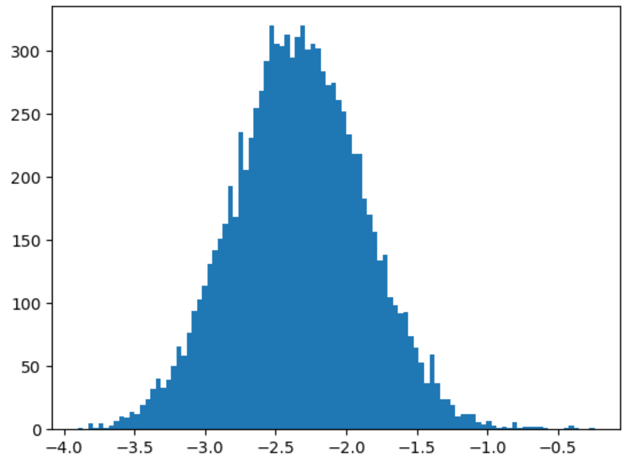
5. **Test a hypothesis of your choice as to which kind of school (e.g. small schools vs. large schools or charter schools vs. not (or any other classification, such as rich vs. poor school)) performs differently than another kind either on some dependent measure, e.g. objective measures of achievement or admission to HSPHS (pick one).**

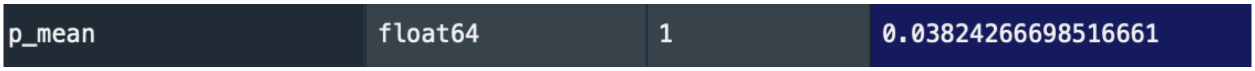
   N0: There is no difference in acceptances to HSPHS between a charter school or not a charter school.

   Na: There is a difference in acceptances to HSPHS between a charter school or not a charter school.

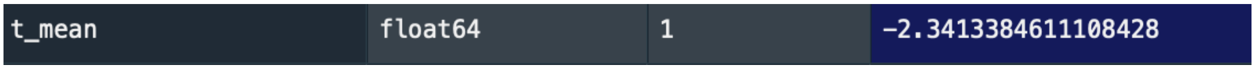
   Based on the data, the schools are categorized into charter schools and non-charter schools. The sample size is set to 50 schools which are drawn randomly from these two independent populations. T-test was used to test the hypothesis as it is independent parametric test. The t-test was done for 10000 repetitions, which shows a normal distribution, and the mean p-value is 0.04 with a t-statistics of -2.34.

   As the p-value = 0.04<0.05, we reject the null hypothesis and conclude that there is a difference in acceptances to HSPHS between a charter school or not a charter school.





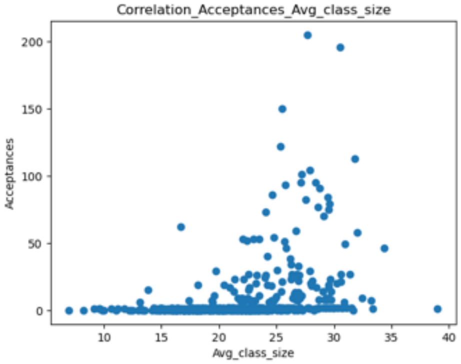
p_mean	float64	1	0.03824266698516661

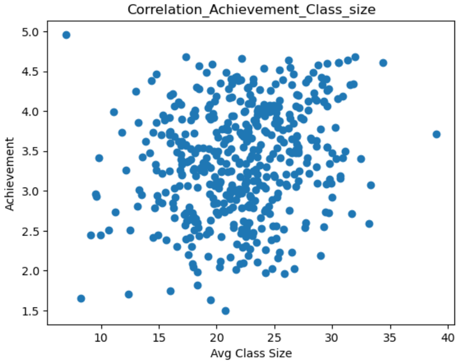


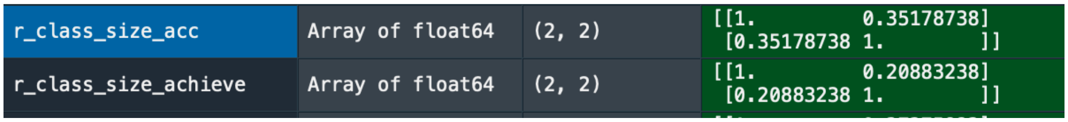
t_mean	float64	1	-2.3413384611108428

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?**

According to the given data, the correlation coefficient between class size and acceptances to HSPHS is 0.35 while the correlation coefficient between class size and student achievement is 0.21. There is not a strong relationship between class size with either measurement. However, the average class size impacts acceptances to HSPHS more than it impacts student achievement, which is also shown by the scatterplots.

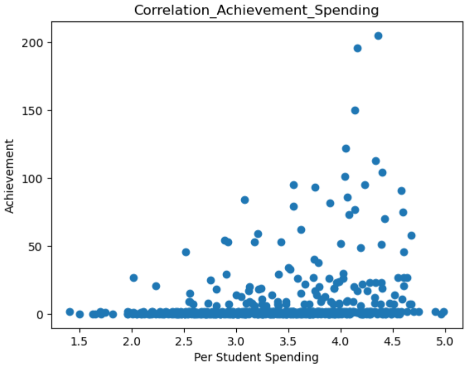


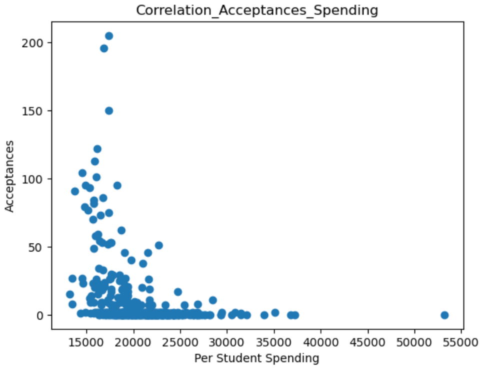


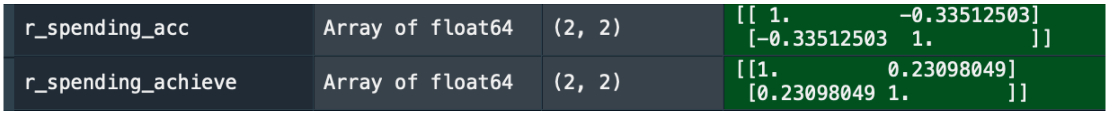


r_class_size_acc	Array of float64	(2, 2)	[[1. 0.35178738] [0.35178738 1.]]
r_class_size_achieve	Array of float64	(2, 2)	[[1. 0.20883238] [0.20883238 1.]]

The correlations between student spending and acceptances to HSPHS or student achievements are also not obvious. The correlation between spending and acceptances to HSPHS is -0.34. As the scatterplot has also shown, there is a negative relationship between these two factors, while the correlation between spending and achievements is 0.23, a positive but not as strongly impacted by the amount of student spending.

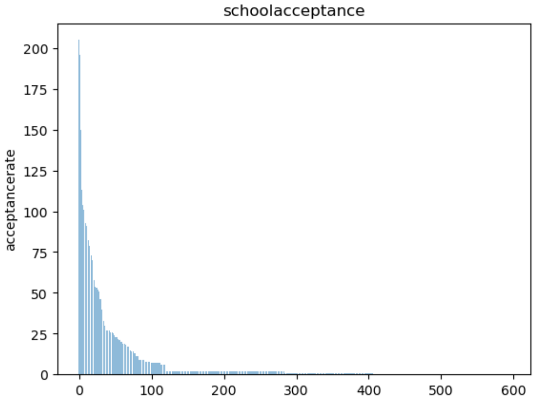


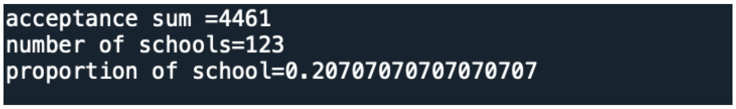




r_spending_acc	Array of float64	(2, 2)	[[1. -0.33512503] [-0.33512503 1.]]
r_spending_achieve	Array of float64	(2, 2)	[[1. 0.23098049] [0.23098049 1.]]

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

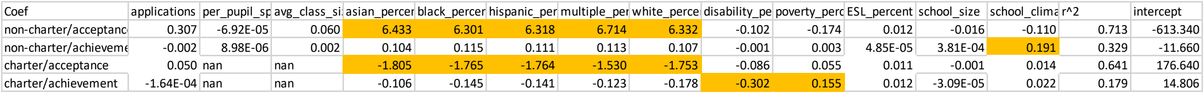




```
acceptance sum =4461
number of schools=123
proportion of school=0.20707070707070707
```

According to the data, the total number of accepted students to HSPHS is 4461 and 90% of the accepted students are 4015. By organizing the number of accepted students in a descending order, the first 123 schools comprise 90% of the students that are accepted by HSPHS. As the total number of schools is 594, the proportion of schools that account for all students accepted to HSPHS is 21%. This result also aligns with the bar graph generated from the data, while the y-axis stands for the number of acceptance and the x-axis is the corresponding schools. The graph is rightly skewed, most of accepted students concentrated in the first 120 schools.

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?**



Coef	applications	per_pupil_sp	avg_class_si	asian_percer	black_percer	hispanic_per	multiple_per	white_perce	disability_pe	poverty_perc	ESL_percent	school_size	school_clima	r^2	intercept
non-charter/acceptanc	0.307	-6.92E-05	0.060	6.433	6.301	6.318	6.714	6.332	-0.102	-0.174	0.012	-0.016	-0.110	0.713	-613.340
non-charter/achieveme	-0.002	8.98E-06	0.002	0.104	0.115	0.111	0.113	0.107	-0.001	0.003	4.85E-05	3.81E-04	0.191	0.329	-11.660
charter/acceptance	0.050	nan	nan	-1.805	-1.765	-1.764	-1.530	-1.753	-0.086	0.055	0.011	-0.001	0.014	0.641	176.640
charter/achievement	-1.64E-04	nan	nan	-0.106	-0.145	-0.141	-0.123	-0.178	-0.302	0.155	0.012	-3.09E-05	0.022	0.179	14.806

As charter schools have systematically missing data, multiple regression has been performed on their specific cases, separated from the non-charter schools by taking out of the E,F columns before cleaning the data by removing NaNs. Factors of how students perceive their schools have gone through the process of dimension reduction and are reduced to the factor 'school_climate'. Objective measures of achievement have also been reduced to a single factor.

By building multiple regression models, the following statistics are generated. From the beta coefficients, it is shown that in non-charter/charter schools, the percent of Asian/Black/Hispanic/Multiple color/White ethnic groups are the most important factors that influence the school's acceptances to HSPHS. In charter schools, percentages of the ethnic groups show a strong negative relationship with the acceptances to HSPHS, while in non-charter schools they influence the acceptances positively. For student achievements, in non-charter schools, the school climate is an important factor that influence it positively while in charter schools, it is the percentage of poverty that affect it positively and the percentage of disability that affect students' achievements negatively, even though the acceptances are barely explained by any of the factors as the r^2 is 0.179.

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

   From the data and analysis of it, different school types, such as a charter school or not a charter school, obviously make a difference in the acceptances of students to HSPHS. In addition, the percentages of different ethnic groups in either type of schools also have impacts on determining the acceptances of students to HSPHS, as for charter schools there is a negative impact and for non-charter schools, the impact is positive. While student spending also exhibits a negative relationship with acceptances, class size shows a positive relationship. However, as analyzed at the beginning of the report, number of applications shows a strong correlation with being accepted by HSPHS with a correlation coefficient of 0.8.

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.**

    In order to send more students to HSPHS, I would recommend encouraging more students to apply and decrease the student spending for both charter schools and non-charter schools. For non-charter schools, specifically, I would recommend increase the student body diversity as it positively impacts the acceptances to HSPHS. For charter schools, I would recommend decrease diversity as it negatively impacts the acceptances.

    By improving objective measures or achievement, I think non-charter schools can focus on improving the school climates, while the charter schools can try to decrease the percentage of disability students and increase the number of students who are under poverty.

Appendix: Code

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Thu Dec 10 22:04:33 2020

@author: moonliang
"""

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats
from sklearn.decomposition import PCA
import random
#1
rawfile = pd.read_csv('middleSchoolData.csv', na_values=["?"])
rawfile_data=rawfile.drop(columns=['dbn','school_name'])
data=rawfile_data.to_numpy()
data=data.astype(np.float)
app=rawfile['applications']
acc=rawfile['acceptances']
plt.scatter(app,acc)
plt.xlabel('app')
plt.ylabel('acc')
plt.show()
corr_app_acc=np.corrcoef(app,acc)

#2
rawfile1=rawfile.drop([250,471])
size=rawfile1['school_size']
app1=rawfile1['applications']
acc1=rawfile1['acceptances']
apprate=(data[:,2])/(data[:,20])
apprate=app1/size
corr_apprate_acc=np.corrcoef(apprate,acc1)
plt.scatter(apprate,acc1)
plt.xlabel('apprate')
plt.ylabel('acceptance')
plt.show()
best=apprate.max()

#3
size=rawfile['school_size']
acc=rawfile['acceptances']
num=rawfile['dbn']
acc_rate=acc/size
best=acc_rate.max()
ind=0
for i in acc_rate:
    if i!=best:
        ind+=1
    else:
        break
print(ind)
plt.scatter(num,acc_rate)
plt.xlabel('schools')
plt.ylabel('acceptancerate')
plt.show()

#4
'''
Q4:  PCA dimensional reduction
'''

student_school=data[:,[9,10,11,12,13,14,19,20,21]]
student_school=student_school.astype(np.float)
a=student_school[~np.isnan(student_school).any(axis=1)]
con_student_school=a[:,[0,1,2,3,4,5]]
con_achieve=a[:,[6,7,8]]


# plt.imshow(con_student_school)
# plt.colorbar()
# r=np.corrcoef(con_student_school,rowvar=False)
# plt.imshow(r)
# plt.colorbar()

zscored_data = stats.zscore(con_student_school)
pca = PCA(n_components=1)
pca.fit(con_student_school)
school_climate=pca.transform(con_student_school)
eig_vals = pca.explained_variance_
loadings = pca.components_
rotated_data = pca.fit_transform(zscored_data)
pca = PCA(n_components=1)
pca.fit(con_achieve)
school_achieve=pca.transform(con_achieve)
r_climate_achieve=np.corrcoef(school_achieve,school_climate,rowvar=False)
plt.scatter(school_achieve,school_climate)
plt.xlabel('school_achieve')
plt.ylabel('school_climate')
plt.show()

covar_explained = eig_vals/sum(eig_vals)*100
num_classes = 6
plt.bar(np.linspace(1,num_classes,num_classes),eig_vals)
#plt.plot(eig_vals)
plt.xlabel('Principal component')
plt.ylabel('Eigenvalue')
plt.plot([0,num_classes],[1,1],color='red',linewidth=1)
```

```
#5
charter=data[485:595]
non_charter=data[0:485]
acceptances_column=3
avg_charter = np.mean(charter[:,acceptances_column])
med_charter = np.median(charter[:,acceptances_column])
std_charter = np.std(charter[:,acceptances_column])
avg_non_charter = np.mean(non_charter[:,acceptances_column])
med_non_charter = np.median(non_charter[:,acceptances_column])
std_non_charter = np.std(non_charter[:,acceptances_column])

num_reps=10000
p_list=[]
t_list=[]
for i in range(num_reps): # loop through each repeat
    sample_charter=np.random.choice(charter[:,3],size=50,replace=False)

sample_non_charter=np.random.choice(non_charter[:,3],size=50,replace=False
)
    t,p=stats.ttest_rel(sample_charter,sample_non_charter)
    p_list.append(p)
    t_list.append(t)
p_mean=np.mean(p_list)
t_mean=np.mean(t_list)

print('p='+str(p_mean),'t='+str(t_mean))
plt.hist(t_list,bins=100)

#6
class_size_acc=data[:,[1,3]]
processed_class_size_acc=class_size_acc[~np.isnan(class_size_acc).any(axis
=1)]
plt.scatter(processed_class_size_acc[:,1],processed_class_size_acc[:,0])
plt.xlabel('Avg_class_size')
plt.ylabel('Acceptances')
plt.title('Correlation_Acceptances_Avg_class_size')
plt.show()
r_class_size_acc =
np.corrcoef(processed_class_size_acc[:,0],processed_class_size_acc[:,1])

spending_acc=data[:,[1,2]]
processed_spending_acc=spending_acc[~np.isnan(spending_acc).any(axis=1)]
plt.scatter(processed_spending_acc[:,1],processed_spending_acc[:,0])
plt.xlabel('Per Student Spending')
plt.ylabel('Acceptances')
plt.title('Correlation_Acceptances_Spending')
plt.show()
r_spending_acc =
np.corrcoef(processed_spending_acc[:,0],processed_spending_acc[:,1])


spending_achieve=data[:,[1,19]]
processed_spending_achieve=spending_achieve[~np.isnan(spending_achieve).an
y(axis=1)]
```

```
plt.scatter(processed_spending_achieve[:,1],processed_spending_achieve[:,0
])
plt.xlabel('Per Student Spending')
plt.ylabel('Achievement')
plt.title('Correlation_Achievement_Spending')
plt.show()
r_spending_achieve =
np.corrcoef(processed_spending_achieve[:,0],processed_spending_achieve[:,1
])

class_size_achieve=data[:,[3,19]]
processed_class_size_achieve=class_size_achieve[~np.isnan(class_size_achie
ve).any(axis=1)]
plt.scatter(processed_class_size_achieve[:,0],processed_class_size_achieve
[:,1])
plt.xlabel('Avg Class Size')
plt.ylabel('Achievement')
plt.title('Correlation_Achievement_Class_size')
plt.show()
r_class_size_achieve =
np.corrcoef(processed_class_size_achieve[:,0],processed_class_size_achieve
[:,1])

#7
import matplotlib.pyplot as plt; plt.rcdefaults()
import numpy as np
import matplotlib.pyplot as plt

schoolnames=rawfile['dbn']
school_acc=dict(zip(schoolnames,acc))
school_accsorted=sorted(school_acc.items(), key=lambda x: x[1],
reverse=True)

objects=[]
values=[]
for i in school_accsorted:
    objects.append(i[0])
    values.append(i[1])
y_pos = np.arange(len(objects))
plt.bar(y_pos,values , align='center', alpha=0.5)
# plt.xticks(y_pos, objects)
plt.ylabel('acceptancerate')
plt.title('schoolacceptance')
plt.show()

sum_acc=acc.sum()
print('acceptance sum ='+ str(sum_acc))
n=0
i=0
while n<=(0.9*sum_acc):
    n+=values[i]
    i+=1
print('number of schools='+str(i))
proportion=123/len(objects)
```

```
print('proportion of school='+str(proportion))

#8
from sklearn import linear_model
'for non-charter schools'
#data1 =
np.transpose([data[:,0],data[:,1],data[:,2],data[:,3],data[:,4],data[:,5],
data[:,6],data[:,7],data[:,8],data[:,9],data[:,10],data[:,11],data[:,12],d
ata[:,13],data[:,14],data[:,15],data[:,16],data[:,17],data[:,18]])
data1=data[~np.isnan(data).any(axis=1)]
student_school1=data1[:,[9,10,11,12,13,14]]
student_school1=student_school1.astype(np.float)
pca = PCA(n_components=1)
pca.fit(student_school1)
school_climate1=pca.transform(student_school1)
data1=np.append(data1,school_climate1,axis=1)
X_noncharter=data1[:,[0,2,3,4,5,6,7,8,15,16,17,18,22]]
Y_acceptance = data1[:,1] # acceptances
regr_a_noncharter = linear_model.LinearRegression() # linearRegression
function from linear_model
regr_a_noncharter.fit(X_noncharter,Y_acceptance) # fit model
print(regr_a_noncharter.score(X_noncharter,Y_acceptance)) # r^2
print(regr_a_noncharter.coef_) # beta
print(regr_a_noncharter.intercept_) # intercept

avg_achieve=data1[:,[19,20,21]]
avg_achieve=avg_achieve.astype(np.float)
pca = PCA(n_components=1)
pca.fit(avg_achieve)
school_achieve1=pca.transform(avg_achieve)
regr_a_noncharter_achieve = linear_model.LinearRegression() #
linearRegression function from linear_model
regr_a_noncharter_achieve.fit(X_noncharter,school_achieve1) # fit model
print(regr_a_noncharter_achieve.score(X_noncharter,school_achieve1)) # r^2
print(regr_a_noncharter_achieve.coef_) # beta
print(regr_a_noncharter_achieve.intercept_)

'for charter schools'
data_charter=data[485:,:]
data2=data_charter[:,[0,1,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21]
]
data2=data2[~np.isnan(data2).any(axis=1)]
student_school2=data2[:,[9,10,11,12,13,14]]
student_school2=student_school2.astype(np.float)
pca = PCA(n_components=1)
pca.fit(student_school2)
school_climate2=pca.transform(student_school2)
data2=np.append(data2,school_climate2,axis=1)
X_charter=data2[:,[0,2,3,4,5,6,7,8,15,16,20]]
Y_acceptance = data2[:,1] # acceptances
regr_a_charter = linear_model.LinearRegression() # linearRegression
function from linear_model
regr_a_charter.fit(X_charter,Y_acceptance) # fit model
print(regr_a_charter.score(X_charter,Y_acceptance)) # r^2
print(regr_a_charter.coef_) # beta
print(regr_a_charter.intercept_) # intercept

avg_achieve2=data2[:,[17,18,19]]
avg_achieve2=avg_achieve2.astype(np.float)
pca = PCA(n_components=1)
pca.fit(avg_achieve2)
school_achieve2=pca.transform(avg_achieve2)
regr_a_charter_achieve = linear_model.LinearRegression() #
linearRegression function from linear_model
regr_a_charter_achieve.fit(X_charter,school_achieve2) # fit model
print(regr_a_charter_achieve.score(X_charter,school_achieve2)) # r^2
print(regr_a_charter_achieve.coef_) # beta
print(regr_a_charter_achieve.intercept_)
```