Getting into New York's Highly Selective Public High Schools (HSPHS)

Analysis of NYC's middle schools and their chances at getting into HSPHS

@author: Saahil Chamdia

This data set is taken from New York City's Department of Education.

In [33]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
import statsmodels.api as sm
from sklearn.preprocessing import StandardScaler
import seaborn as sns
from scipy import stats
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn import metrics
pd.options.mode.chained_assignment = None
```

In [2]:
```
data = pd.read_csv('middleSchoolData.csv')
```

Cleaning the Data

Initially, imputing the median value (column-wise) for all missing values was considered. However, the imputed values could be misrepresent the truth- as the real value could vary significantly- and it could have an impact on the principal component analysis that would be conducted later on. At the same time, dropping all schools with any null fields could greatly reduce the power of our study. Therefore, to minimize the effect of imputation and maximize power, **a threshold of 20 null values** was decided upon for which values would be imupted.

In [3]:
```
# Dropping all NaN values for columns that have more than 20 NaNs
data.dropna(thresh=20, inplace=True)
# Filling NaN values with the column median
data = data.fillna(data.median())
print(data.shape)
```

```
(544, 24)
```

After cleaning the data (dropping certian schools due to missing data and imputing certain values), we have workable data on 544 schools. Hence, due to using a threshold, our sample size has not reduced significantly.

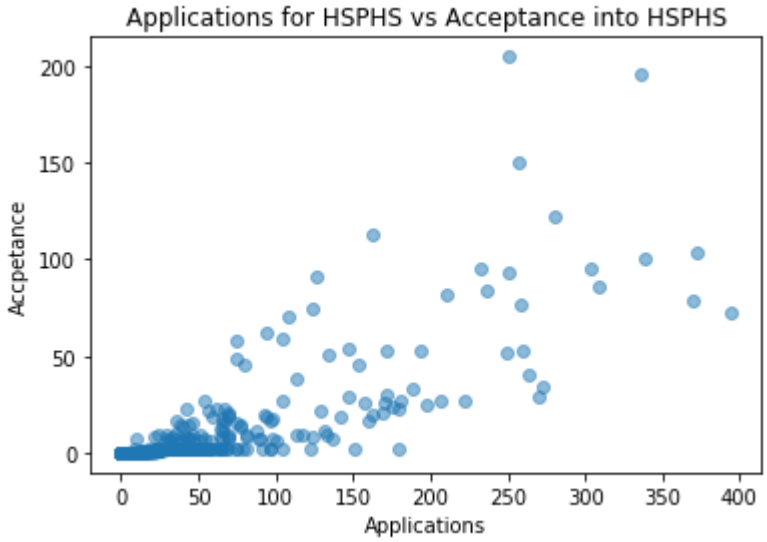
In [28]:
```
# creating a new dataframe that only contains numerical data. This will be use
d in multiple cases.
data_nums = data.drop('dbn', axis = 'columns')
data_nums.drop('school_name', axis = 'columns', inplace=True)
```

Analysis

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

In [76]:
```
plt.scatter(data['applications'], data["acceptances"], alpha = 0.5)
plt.xlabel("Applications")
plt.ylabel("Accpetance")
plt.title("Applications for HSPHS vs Acceptance into HSPHS")
```

Out[76]: Text(0.5, 1.0, 'Applications for HSPHS vs Acceptance into HSPHS')



In [132]:
```
np.corrcoef(data['applications'], data["acceptances"])[0][1]
```

Out[132]: 0.8026638536729013

The correlation between applications and admissions into HSPHS is 0.803.

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

The raw number of applications gives us no indication of the proportion of students from a particular middle school that apply for HSPHS. For example, when looking at raw application scores, East Side Community School (ESCS) has 16 while the School for Global Leaders (SGL) has 19. However, the application rate to HSPHS for ESCS is much higher than that of the SGL as the average class size at ESCS is almost half (10.64) of SGL's (20.33). Hence, **application rate is a better predictor of admissions** into HSPHS than the raw number of applications.

3) Which school has the best *per student* odds of sending someone to HSPHS?
(note: The use of 'school_size' means that all students in the middle school are factored into the following calculation, not just the students graduating to high school)

In [78]:
```
odds = data['acceptances']/data['school_size']
max_odds = max(odds)
max_odds_index = odds.idxmax()
print(data.school_name[304])
print(round(max_odds, 3))
```

```
THE CHRISTA MCAULIFFE SCHOOL\I.S. 187
0.235
```

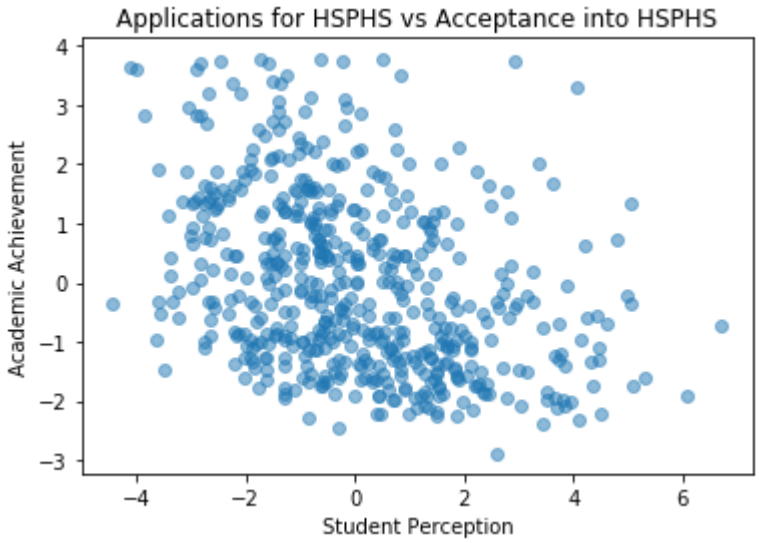
THE CHRISTA MCAULIFFE SCHOOL\I.S. 187 has the best odds of sending one of their students to a HSPHS **odds** $\approx$ **1 : 3** (when rounding the probability to 0.25).

4) Is there a relationship between how students perceive their school and how the school performs on objective measures of achievement?

To answer this question, the dimension of columns L-Q (which consist of responses that represent the students' perception of their school) and the dimensions of columns V-X (which consists of objective academic measures) will be reduced to 1 each using PCA. This will make it easier to deduce the correlation between the students' perception of their school and the academic achievement of the school.

In [5]:
```
# Columns L-Q
data_perception = data_nums.copy()
data_perception.drop(data_perception.columns[[0,1,2,3,4,5,6,7,8,15,16,17,18,19
,20,21]], axis = 1, inplace = True)
# Columns V-X
data_achievement = data_nums.copy()
data_achievement.drop(data_achievement.columns[[0,1,2,3,4,5,6,7,8,9,10,11,12,1
3,14,15,16,17,18]], axis = 1, inplace = True)
#Normalizing the data
features1 = ['rigorous_instruction', 'collaborative_teachers', 'supportive_env
ironment', 'effective_school_leadership', 'strong_family_community_ties', 'tru
st']
features2 = ['student_achievement', 'reading_scores_exceed', 'math_scores_exce
ed']
x1 = data_perception.loc[:, features1].values
x1 = StandardScaler().fit_transform(x1)
x2 = data_achievement.loc[:, features2].values
x2 = StandardScaler().fit_transform(x2)
# Reducing dimension to 1 for each
pca = PCA(n_components=1)
principalComponents1 = pca.fit_transform(x1)
principalComponents2 = pca.fit_transform(x2)
principalDf1 = pd.DataFrame(data = principalComponents1, columns = ['student_p
erception'])
principalDf2 = pd.DataFrame(data = principalComponents2, columns = ['academic_
achievement'])
# Combining the results into one dataframe
pca_result = pd.concat([principalDf1, principalDf2], axis=1)
# Plotting
plt.scatter(pca_result['student_perception'], pca_result["academic_achievemen
t"], alpha = 0.5)
plt.xlabel("Student Perception")
plt.ylabel("Academic Achievement")
plt.title("Applications for HSPHS vs Acceptance into HSPHS")
```

Out[5]: Text(0.5, 1.0, 'Applications for HSPHS vs Acceptance into HSPHS')



In [80]:
```
np.corrcoef(pca_result['student_perception'], pca_result["academic_achievement"])[0][1]
```

Out[80]:
```
-0.37008261361034944
```

Therefore, **there seems to be a negative correlation** between the students' perception of their school and the academic achievement of their school. This means that generally, the higher the perception a student has of their school, the worse than school performs academically. However, as this is a simple correlation, we are not establishing a link between the two factors.

5) Hypothesis Test to test the relation between class size and a school's academic excellency.
(Academic excellency is measured as the proportion of students exceeding state-wide expectations in reading and math)

H_o: There is no effect of aving small class sizes on a school's academic peformance
H_a: Small class room sizes have an effect on the school's academic peformance

Treatment assignment process: schools which average classes that are smaller than the median are considered as treated.
Test Type: Z Test

In [22]:
```
# create a copy of the original dataset as we will be adding a new column
data_size = data[['avg_class_size', 'math_scores_exceed', 'reading_scores_exceed']].copy().reset_index()
# add a new column that holds the gerenal academic excellency of students at a particular school
data_size['culmulative_scores'] = data_size[['math_scores_exceed', 'reading_scores_exceed']].mean(axis=1)
size_boundary = data['avg_class_size'].median()
# Treatment Assignment
data_size['treatment'] = 1
for i in range(len(data_size)):
    if (data_size['avg_class_size'][i] >= size_boundary):
        data_size['treatment'][i] = 0
# Difference in Means Estimator
control = data_size[data_size["treatment"] == 0]
treatment = data_size[data_size["treatment"] == 1]
mean_diff = print(treatment["culmulative_scores"].mean() - control["culmulative_scores"].mean())
print(mean_diff)
```

```
-0.1945519670923722
None
```

Therefore, looking at the average treatment effect using a difference in means estimator, there seems to be a **negative treatment effect** on the outcome. This means that smaller class sizes lead to lower academic excellency. Now we test if this finding is statistically significant at an alpha level of 0.05:

In [23]:
```
# Z-Test
x = np.concatenate((treatment["culmulative_scores"], control["culmulative_scor
es"]))
k2, p = stats.normaltest(x)
print("p-value = {:g}".format(p))
```

```
p-value = 4.77106e-11
```

Therefore, at a signficance level of 0.05, we have **sufficient evidence to reject the null** that claimed that small classroom sizes have no effect on the academic excellency of middle schools in NYC. Our negative estimate inisuates that smaller classrooms lead to a lower aacademic excellency of a school.

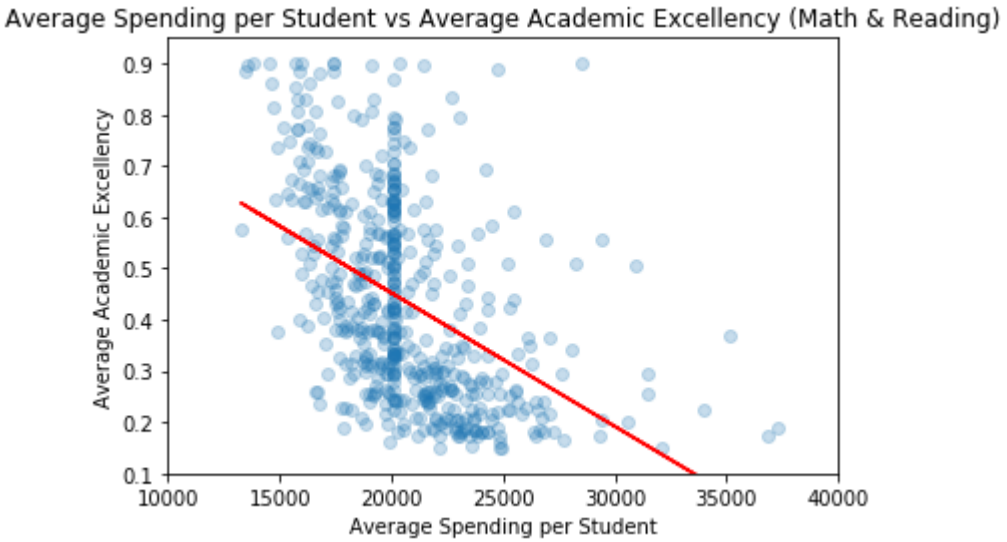
6) Is there any evidence that the availability of material resources impacts objective measures of achievement or admission to HSPHS?

To assess this, a regression analysis between the average spending per student in a given school and that school's academic excellency (Proportion of students exceeding state-wide expectations in reading and math). This is done too see if there is a relation between the average money students spend at a school and the academic excellency of that school.

In [26]:
```
data_resources = data.copy()
data_resources['culmulative_scores'] = data_resources[['reading_scores_exceed'
, 'math_scores_exceed']].mean(axis=1)
# Creating the Regression Model
x = data_resources["per_pupil_spending"]
y = data_resources['culmulative_scores']
X = sm.add_constant(x)
model = sm.OLS(y, X)
results = model.fit()
alpha = results.params[0]
beta = results.params[1]
print("Regression Equation: \n(y = Average Cumulative Scores, x = Average Stud
ent Spending)")
print("\ny = (",beta,"* x)", alpha)
# Plotting our findings
plt.xlim(10000, 40000)
plt.ylim(0.1, .95)
plt.scatter(data_resources['per_pupil_spending'], data_resources["culmulative_
scores"], alpha = 0.25)
plt.plot(x,results.params[0] + x*results.params[1], color = "red", label = 'y
 = 0.33x - 2.92')
plt.xlabel("Average Spending per Student")
plt.ylabel("Average Academic Excellency")
plt.title("Average Spending per Student vs Average Academic Excellency (Math &
Reading)")
plt.show()
```

```
Regression Equation:
(y = Average Cumulative Scores, x = Average Student Spending)

y = ( -2.6046948715836496e-05 * x) 0.9733045064763529
```



In [48]:
```
results.summary()
```

Out[48]:

OLS Regression Results

Dep. Variable:	culmulative_scores	**R-squared:**	0.225
Model:	OLS	**Adj. R-squared:**	0.224
Method:	Least Squares	**F-statistic:**	157.6
Date:	Mon, 07 Dec 2020	**Prob (F-statistic):**	6.41e-32
Time:	23:40:57	**Log-Likelihood:**	181.59
No. Observations:	544	**AIC:**	-359.2
Df Residuals:	542	**BIC:**	-350.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.9733	0.043	22.401	0.000	0.888	1.059
per_pupil_spending	-2.605e-05	2.07e-06	-12.556	0.000	-3.01e-05	-2.2e-05

Omnibus:	41.262	**Durbin-Watson:**	1.486
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	48.837
Skew:	0.727	**Prob(JB):**	2.48e-11
Kurtosis:	3.204	**Cond. No.**	1.22e+05

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.22e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

At a significance level of 0.05, we have sufficient evidence to reject the null hypothesis that claims that there is no effect of average spending per student at middle schools on the schools academic excellency. We observe a negative coefficient, insinuating a negative relation between the average spending of a student and their's school's academic excellency, meaning that the higher a student spends on average, the more likely that their school would have a lower proportion of students that exceed state-wide expectations.

While the plotted line suggests a negative, linear trend, it must be noted that the value for R^2 (0.225) is considerably low. This implies that the variance in the average cumulative scores for each school is not explained well by the average money students at the school spend.

Therefore, **while the regression model insinuates that there is a negative, linear trend between the stated variables, the average money spent by students at a school is not a good predictor of the school's academic excellency.** Hence, with the given data, we cannot effectively claim that access to more learning resources (indicated by a higher average student spending) has a impactful effect on the school's academic excellency.

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

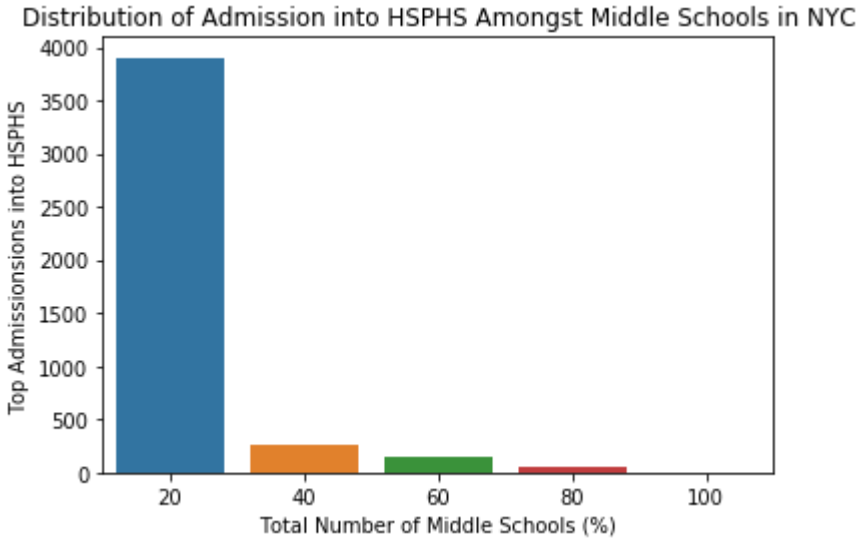
In [12]:
```
data_admin = data.copy()
# Number of admissions that make 90% of all admissions:
admin_limit = int(0.9 * sum(data_admin.acceptances))
# sorting the data in descending value
data_admin.sort_values(by=['acceptances'], ascending = False, inplace = True,
ignore_index = True)
sum_accepted = 0
index = 0
# Counts how many schools make up 90% of all admissions into HSPHS
for i in range(len(data_admin)):
    sum_accepted = sum_accepted + data_admin['acceptances'][i]
    if sum_accepted >= admin_limit:
        index = i
        break
print(index/len(data_admin))
```

```
0.20220588235294118
```

Hence, **around 20.2%** of all middle schools in NYC make up for 90% of the admissions into HSPHS.
(note: few schools were dropped initally due to missing data. Thus, the proportion of admissions might vary slightly if one were to include those schools too)

Whithin these schools, here is the distribution of admission:

In [62]:
```
# Assignment of quantiles
quantile = [int((1/5)*len(data_admin)),int((2/5)*len(data_admin)),int((3/5)*le
n(data_admin)), int((4/5)*len(data_admin))]
data_admin['quantile'] = 100
for i in range(len(data_admin)):
    if(i <= quantile[3]):
        data_admin['quantile'][i] = 80
    if(i <= quantile[2]):
        data_admin['quantile'][i] = 60
    if(i <= quantile[1]):
        data_admin['quantile'][i] = 40
    if(i <= quantile[0]):
        data_admin['quantile'][i] = 20
# Plotting the distribution
plt.figure(num=None, dpi=80, facecolor='w', edgecolor='r')
data_admin_agg = data_admin.groupby(['quantile']).sum().reset_index()
ax = sns.barplot(x="quantile", y = "acceptances" , data = data_admin_agg)
plt.ylabel("Top Admissionsions into HSPHS")
plt.xlabel("Total Number of Middle Schools (%)")
plt.title("Distribution of Admission into HSPHS Amongst Middle Schools in NYC"
)
plt.show()
```

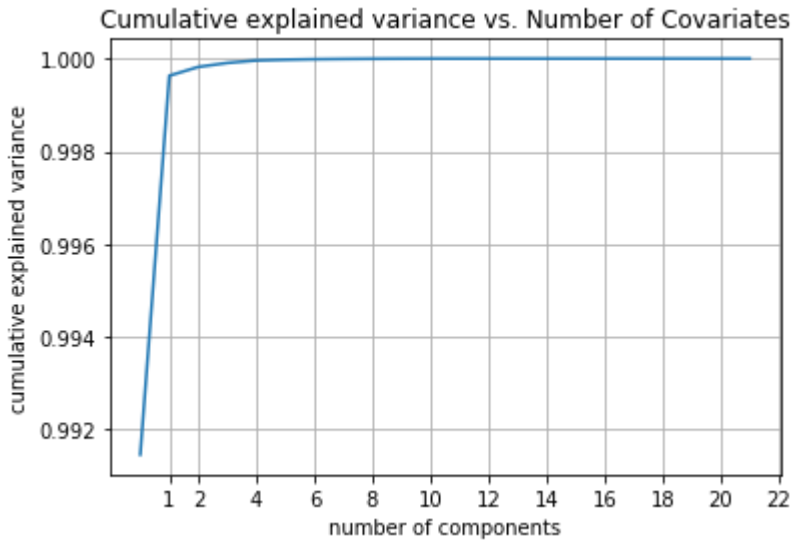


The figure above illustrates how around 20% of middle schools in NYC contribute to about 90% of admissions into HSPHS.

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?

In [27]:
```
# Plotting the cumulative explained variance to determine the required covaria
tes
pca = PCA().fit(data_nums)
plt.plot(np.cumsum(pca.explained_variance_ratio_))
plt.xticks([1,2,4,6,8,10,12,14,16,18,20,22])
plt.xlabel('number of components')
plt.ylabel('cumulative explained variance')
plt.grid()
plt.title("Cumulative explained variance vs. Number of Covariates")
```

Out[27]:
```
Text(0.5, 1.0, 'Cumulative explained variance vs. Number of Covariates')
```



As seen in the figure above, we could explain almost a 100% of the variance in our dataset with only 4 features. This is useful as it gives us an insight into the number of factors that the most important in understanding our dataset.

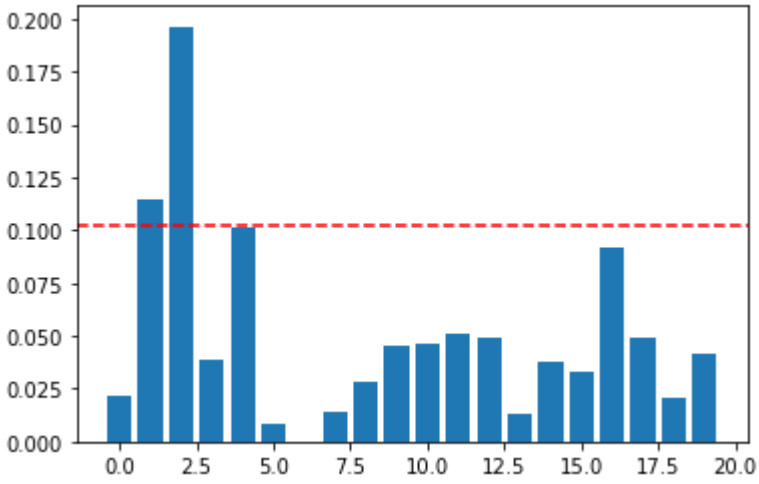
a) Factors that are important for a school to have a high admission proportion for HSPHS:
(A decision Tree will be used to assess the importance of all features)

In [46]:
```
data_admission = data_nums.copy().reset_index()
data_admission['admission_proportion'] = data_admission['acceptances']/data_admission['school_size']
# Binary Indicator of whether a school has a high aceptance proportion based on median admission proportion
data_admission['outcome'] = 0
for i in range(len(data_admission)):
    if(data_admission['admission_proportion'][i]>=data_admission['admission_proportion'].median()):
        data_admission['outcome'][i] = 1
feature_cols = ['per_pupil_spending', 'avg_class_size', 'asian_percent', 'black_percent','hispanic_percent','multiple_percent','multiple_percent', 'rigorous_instruction','collaborative_teachers','supportive_environment', 'effective_school_leadership','strong_family_community_ties', 'trust','disability_percent', 'poverty_percent','ESL_percent','school_size','student_achievement','reading_scores_exceed','math_scores_exceed']
# The features
X = data_admission[feature_cols]
# The target Variable (proportion of admissions)
y = data_admission.outcome
# Training the Model
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1) # 70% training and 30% test
clf = DecisionTreeClassifier()
# Train Decision Tree Classifer
clf = clf.fit(X_train,y_train)
#Predict the response for test dataset
y_pred = clf.predict(X_test)
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.676829268292683
```

The accuracy of our Decision Tree Model is ~70% which is considered relaible. Hence, we can effectively use our Decision tree to draw feature importances to understand what features contribute the greatest to the admission proportion into HSPHS (Using the 3rd Criterion- Factors that explain 90% of the variance):

In [72]:
```
# Importance of each feature
importance = clf.feature_importances_
important_features = []
for i,v in enumerate(importance):
    # Any feature that is of the 90th perectile is considered important
    if (v >= np.quantile(importance, 0.89)):
        important_features.append(i)
# plot feature importance and cutoff
plt.bar([x for x in range(len(importance))], importance)
plt.axhline(y=np.quantile(importance, 0.9), color='r', linestyle='--')
plt.show()
print('\033[1m', "Important Features:", "\033[0;0m")
for i in important_features:
    print(" ",feature_cols[i])
```




```
Important Features:
 avg_class_size
 asian_percent
 hispanic_percent
```

Therefore, 3 recorded variables contributed the most towards a shcool's chance of having a higher admissions into HSPHS proportion: average class size, the percentage of asian students, the percentage of hispanic students. This provides valuable insight as it tells us that larger classess and a specfic racial demographic can be predictive of sending students to HSPHS.

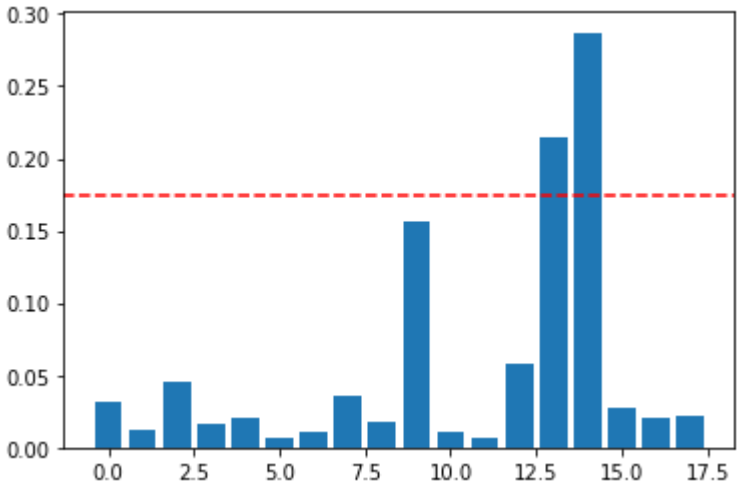
b) achieving high scores on objective measures of achievement?
(A decision Tree will be used to assess the importance of all features)

In [81]:
```
# Binary Indicator of whether a school has high academic excellency
data_admission['outcome'] = 0
data_admission['cumulative_scores'] = data_admission[['reading_scores_exceed',
'math_scores_exceed']].mean(axis=1)
for i in range(len(data_admission)):
    if(data_admission['cumulative_scores'][i]>=data_admission['cumulative_scor
es'].median()):
        data_admission['outcome'][i] = 1
feature_cols = ['per_pupil_spending', 'avg_class_size', 'asian_percent', 'blac
k_percent','hispanic_percent','multiple_percent','multiple_percent', 'rigorous
_instruction','collaborative_teachers','supportive_environment', 'effective_sc
hool_leadership','strong_family_community_ties', 'trust','disability_percent',
'poverty_percent','ESL_percent','school_size','student_achievement']
# The features
X = data_admission[feature_cols]
# The target Variable (high academic excellency)
y = data_admission.outcome
# Training the Model
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, rando
m_state=1) # 70% training and 30% test
clf = DecisionTreeClassifier()
# Train Decision Tree Classifer
clf = clf.fit(X_train,y_train)
#Predict the response for test dataset
y_pred = clf.predict(X_test)
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.7560975609756098
```

The accuracy of our Decision Tree Model is ~75% which is considered relaible. Hence, we can effectively use our Decision tree to draw feature importances to understand what features contribute the greatest to the objective achievement of a school (Using the 3rd Criterion- Factors that explain 90% of the variance):

In [80]:
```
# Importance of each feature
importance = clf.feature_importances_
important_features = []
for i,v in enumerate(importance):
    # Any feature that is of the 90th perectile is considered important
    if (v >= np.quantile(importance, 0.90)):
        important_features.append(i)
# plot feature importance and cutoff
plt.bar([x for x in range(len(importance))], importance)
plt.axhline(y=np.quantile(importance, 0.9), color='r', linestyle='--')
plt.show()
print('\033[1m', "Important Features:", "\033[0;0m")
for i in important_features:
    print(" ",feature_cols[i])
```



Important Features:
 disability_percent
 poverty_percent

Therefore, 2 recorded variables contributed the most towards a shcool's objective academic achievement: the percentage of disabled kids and the percentage of poor kids at a school. This provides valuable insight as it tells us that the composition of a school based on income and disability is predictive of the establishment's academic standing.

Discussion

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 question 5, we observed that smaller classes seemed to have a statistically significant, negative correlation with the average academic excellency of a middle shcool. However, as we did not control for confouders, we cannot establish a causal link between the two. Similarly, we observed that a higher avergae student expenditure was also negatively correlated to academic excellence. However, as the coefficient of determination for that statistical analysis was small, we stated that student expenditure was not a good predictor of academic excellence.

Using our first Decision Tree Model, it was evident that racial demographic plays a crucial role on the proportion of students sent to HSPHS from middle schools, expecially members of the asian and hispanic community. Moreover, the average class sizes also played a role, suggesting that our findings in question 5 might be indicative of a strong predictive link. Using our second decision tree model, we learnt that the percetage of disabled students and the percentage of students in poverty are indicative of the schools academic performance. As academic peformance is important for admission into HSPHS, it is important to note that these factors might also indirectly affect the chances a middle school has for sending its students to HSPHS.

10) Imagine that you are working for the New York City Department of Education as a data scientist. 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.

As we learnt that racial demographic plays a big role on admission into HSPHS, I would reccomend the Department of Education to promote racial diversity in middle schools. Balancing middle schools on demograhpics would help control . This would definitely be an ordeal as a lot of factors like location, fees, academics etc. are considered when parents enroll their children at a particular school but incentivising diverse admissions and providing better transportation options could help bridge this inbalance. Additionally, class sizes also play a role on selection. We also found out that smaller class sizes are negatively correlated to academic performance. Hence, I would recommend insisting schools to aim for a standardized range for their class size.

For improvement to objective measures of academic excellence, I would recommend increasing the Deparment of Education's academic/financial support for middle schools that have a larger proportion of disabled students and students in poverty. Our decision tree model indicated that these two factors have the biggest impact on the academic standing of a middle school and therefore, if the city focuses on providing more resources to middleschools with this demographic, students that fit this category might have a better chance at getting admitted to a HSPHS, thus lessening the disparity of thier demographics in HSPHS.

In [ ]: