Introduction to Data Science (DS-UA 112) Fall 2020 Final Project

Erin Choi

In the 'middleSchoolData.csv' dataset, the NYC Department of Education provides information about the characteristics of 594 NYC middle schools, such as factors that describe school climate and measures of student achievement, as well as the numbers of applications and acceptances to one of 8 highly selective public high schools (HSPHS).

Dimension Reduction and Cleaning the Data

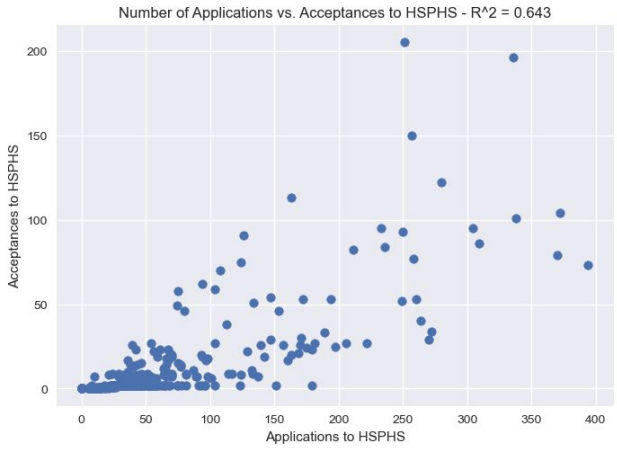
PCA was performed for Questions 4 and 8, as these questions involved an analysis of the effects of multiple factors on another group of factors at once. For other questions, which looked at smaller numbers of specific factors, variable selection was applied. A copy of the original dataset was made, and only the columns with factors relevant to specific questions were retained before performing analyses in order to make the data easier to read and slice.

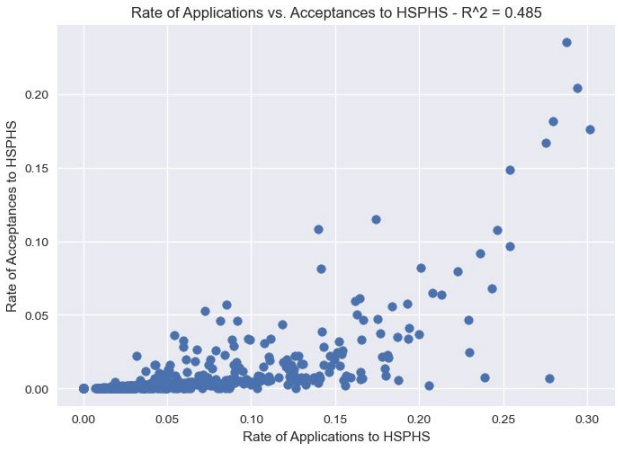
Data was cleaned on a question-by-question basis in order to retain as much data as possible depending on each question's factors of interest. Rows with missing data were removed after dimension reduction/selection. Removing all rows with missing data at the beginning would remove schools that could be used for some analyses but not others. The only question for which all rows with NaN values were removed was Question 8, as it involves all factors. (All charter schools are missing per pupil spending and average class size data, so they were excluded from the construction of the models in Question 8.)

Questions:

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

The correlation between the number of applications and admissions to HSPHS was 0.802. This matches the scatterplot of the data, shown below on the left, which displays a moderate to strong positive correlation. It also makes sense given that admission to HSPHS is contingent on applying. The data appears to be skewed to the right, with most schools having relatively lower numbers of applications and acceptances to HSPHS.

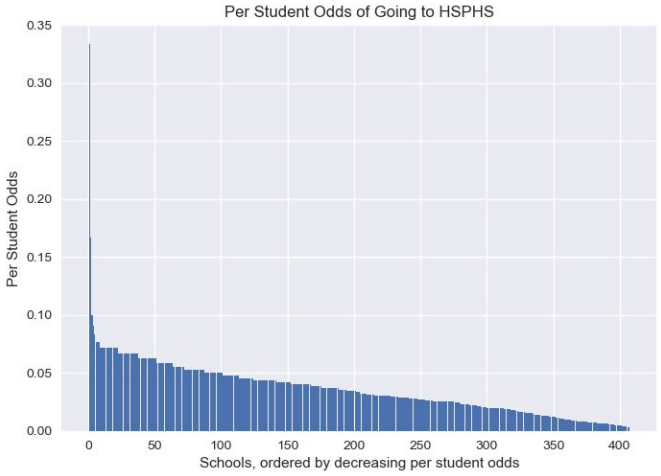




2. What is a better predictor of admission to HSPHS: the raw number of applications or application *rate*?

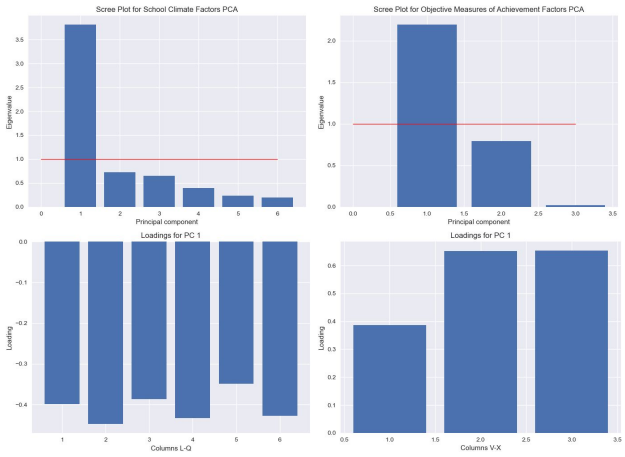
   The R^2 value of the relationship between the number of applications and acceptances to HSPHS is 0.643. Meanwhile, the R^2 of the relationship between the application rate and acceptance rate to HSPHS is 0.485 (scatterplot shown above on the right). Since the R^2 value of the former relationship is greater, the raw number of applications appears to be a better predictor of admission to HSPHS; this value can be interpreted to mean the change in number of applications explains 64.3% of the variation in the number of acceptances to HSPHS.

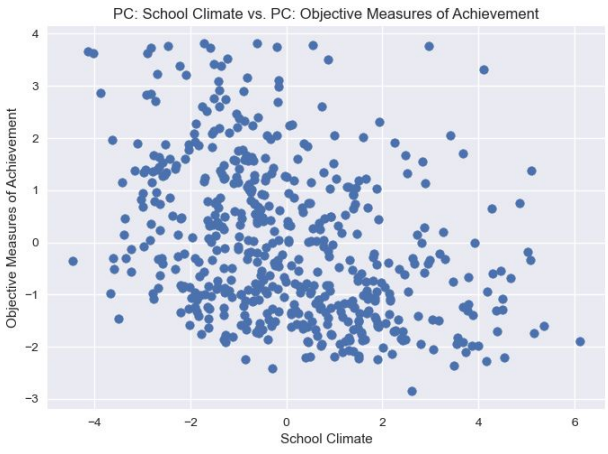
3. Which school has the best *per student* odds of sending someone to HSPHS?



The odds of acceptance to HSPHS at each school was computed as the ratio of accepted students to applicants who were not accepted. The per student odds for each school were found by normalizing the odds by the number of acceptances. Based on these calculations, the school with the best per student odds was determined to be the Special Music School, with a per student odds value of 0.333. The bar chart to the left, which displays the per student odds of all schools in decreasing order, shows that this particular school has a much greater per student odds value than other schools.

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)?

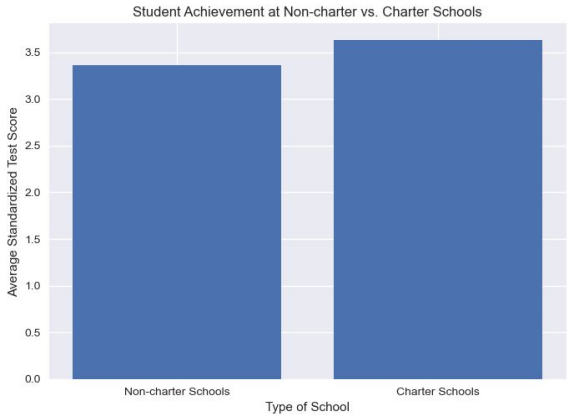




Two PCAs were performed to reduce the dimensions of this dataset into two factors: one on columns L-Q and another on columns V-X. The scree plot with the Kaiser criterion line and loadings chart for the first principal component of each PCA are shown above (the charts using

columns L-Q are on the left of those using columns V-X). The two first principal components, one from each PCA, were correlated, yielding a correlation of -0.367. Based on the this value and the scatterplot of this relationship, shown above on the right, there appears to be a considerably weak negative relationship between how students perceive their school (or "school climate"), such as measures of how rigorous instruction is, and objective measures of achievement, including average scores on a state-wide standardized test.

5. Test a hypothesis of your choice as to which kind of school performs differently than another kind either on some dependent measure or admission to HSPHS.



I asked the question, "is there a difference between the average standardized test scores of charter schools and non-charter schools?" The null hypothesis is that there is no difference between the average standardized test scores of charter and non-charter schools. After splitting the schools into two groups based on whether they are a charter school or not, I performed an independent samples t-test on the two groups, with the dependent variable being the 'student_achievement' column in the dataset. The test yielded a t-value of -3.071 and a p-value of 0.002. This p-value is less than 0.01, so I rejected the null hypothesis at the α=0.01 level and concluded that there is enough evidence to say there *is* a difference between the average standardized test scores of charter and non-charter schools. The bar graph above shows the mean average standardized test scores of each type of school.

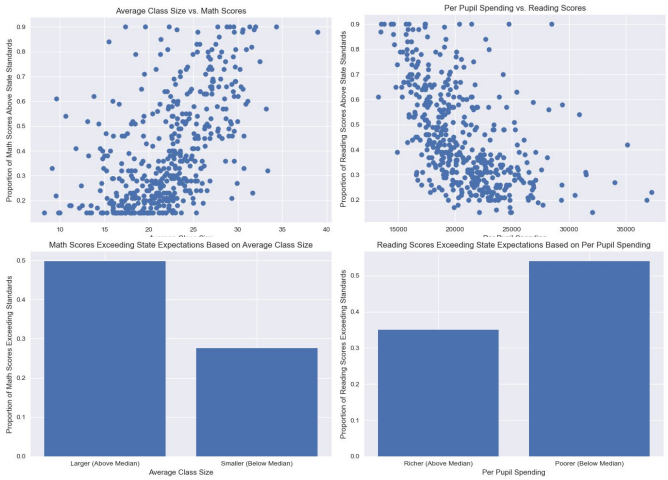
6. Is there any evidence that the availability of material resources (e.g. per-student spending or class size) impacts objective measures of achievement or admission to HSPHS?

Correlations between Material Resource Factors and Admission/Achievement Factors

	Per pupil spending	Average class size
Acceptances	-0.356	0.350
Acceptance rate	-0.307	0.348
Student achievement	-0.158	0.209
Reading scores exceed	-0.498	0.537
Math scores exceed	-0.485	0.558

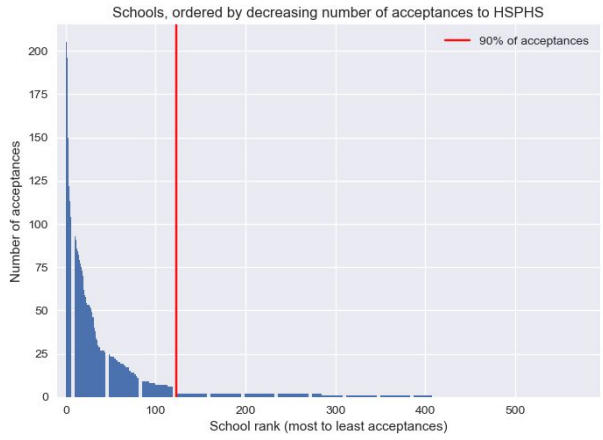
I investigated the correlations between the two material resource factors and the achievement and admission factors individually; the correlations are displayed in the table above. All objective measures of achievement are negatively correlated with spending per student and positively correlated with average class size. One would think more spending along with smaller class sizes

leads to better results in achievement, but both factors point in the opposite direction. I further investigated the factors that yielded the greatest correlations in each column—average class size with math scores and per pupil spending with reading scores (the scatterplots that illustrate these correlations are displayed below)—by performing two independent samples t-tests.



I transformed average class size and per pupil spending into categorical variables—they became "larger" and "smaller" class size and "richer" and "poorer" schools, respectively—by dividing each factor at its median. The means of each group are displayed in the bar graphs, with the graph for class size to the left of the graph for spending. The test between the different class sizes resulted in a t-value of 13.05 and a p-value that was essentially zero (1.898e-33). I concluded that there is enough evidence to say there is a significant difference between the proportions of math scores exceeding state standards at schools with higher versus lower class sizes. The test between the richer and poorer schools gave me a t-value of -12.31 and a p-value of basically zero as well (2.138e-30); there is enough evidence to say there is a significant difference between the proportions of reading scores above state standards at richer versus poorer schools. The results of the two tests are evidence that objective measures of achievement are different at schools with different amounts of material resources available.

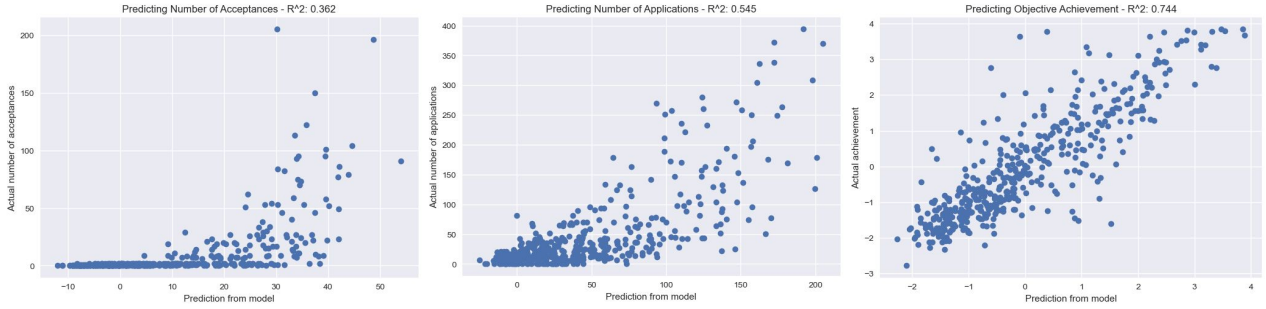
7. What proportion of schools accounts for 90% of all students accepted to HSPHS?



I summed up the acceptances from all schools and determined a "threshold" at 90% of that sum. After sorting the schools from most to least acceptances, I found the index of the school where the cumulative sum of acceptances passed the threshold and divided the index+1 by the total number of schools. The proportion of schools that accounts for 90% of all acceptances is 0.207; in other words, 20.7% of the schools account for 90% of all students accepted to HSPHS. The bar graph to the left shows all schools ordered by decreasing number of acceptances, with the schools to the left of the red line being those that account for 90% of acceptances to HSPHS.

8. Build a model that includes all factors as to what school characteristics are most important in terms of a) sending students to HSPHS and b) achieving high scores on objective measures of achievement.

   After performing multiple PCAs to reduce many factors to fewer representative factors, I built three multiple regression models that predicted number of acceptances, number of applications, and objective measures of achievement (as one factor reduced by PCA). I included the model for not only acceptances but also applications, since admission is contingent on applying. Displayed below are each models' predictions plotted against the actual values of the dependent factor. All three models used the same 7 predictors, which are shown with their betas for each model in the table beneath the scatterplots.



Betas (coefficients) of Predictors for Each Multiple Regression Model

	Per pupil spending	Average class size	Diversity AMW (Asian, Multi, White - col. G, J-K)	Diversity BH (Black, Hispanic - col. H-I)	School climate (col. L-Q)	Disadvantaged students (col. R-T)	School size
Acceptances	-1.866e-05	0.286764	6.21205	-0.044288	-0.381451	1.5958	0.017751
Applications	0.000265	0.625493	10.4107	1.9178	-1.16955	-0.401351	0.109283
Objective Achievement (col. V-X)	-1.220e-05	0.024839	0.435727	-0.211348	-0.167071	0.487147	0.000187

* Factors combined/reduced by PCA are highlighted in blue.
** Numbers have been rounded to 6 decimal places.

In both the model for acceptances and for applications, the diversity AMW (percentages of Asian, multiracial, and white students) factor had the highest beta. The objective achievement model's highest beta was for the disadvantaged students (percentages of students who are disabled, in poverty, or ESL), with the diversity AMW factor being the second highest. According to these models, it appears that the percentages of white/multiracial/Asian students and disadvantaged students are the most important in predicting how many students will get accepted to HSPHS and performance in objective measures of achievement. Factors that were investigated in previous questions, such as school climate and spending, held less weight in the models, so while they may have relevant relationships with the dependent factors, they seem to be less important in predicting acceptances and achievement.

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

   The strong positive correlation between applications and acceptances supports the contingency of acceptance on applying. There appears to be a weak negative relationship between students' perception of their schools, or school climate, and objective measures of achievement. There is also evidence of the impact of material resource availability on achievement, but the relationship between these factors is the opposite of what is expected; measures of achievement are negatively correlated with spending and positively correlated with class size. Furthermore, there is evidence of a significant difference between the standardized test scores of charter schools and non-charter schools. While these relationships are relevant, multiple regression involving all factors reveals that high percentages of white, Asian, and multiracial students as well as high percentages of disadvantaged students are seemingly most important in determining acceptances to HSPHS and performance in objective measures of achievement.

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 of achievement?

    It appears that the percentage of Asian, multiracial, and white students at a school plays an important part in predicting acceptances to HSPHS and achievement. These groups are often considered privileged, particularly in comparison to the other groups included in the data (Black and Hispanic). The percentage of disabled, impoverished, and/or ESL students is also important; it is worth noting that in question 8, this factor has a positive beta in the models for acceptances and achievement but a negative beta for the model for applications. This may suggest disadvantaged students face difficulties applying to HSPHS despite good academic performance, and again, admission is contingent on applying.

    I would recommend that the NYC Department of Education advocate for more diversity in NYC middle schools in terms of ethnic background while increasing support for underprivileged Black and Hispanic students. These students may not have access to resources that more privileged students have, especially outside of school, so the Department of Education should strive to make up for such a lack of external opportunities by providing more opportunities within the schools themselves, perhaps via development or expansion of clubs or programs with an academic focus. I would also recommend providing the appropriate resources for schools to better support disadvantaged students during the HSPHS application process. There may be talented students who encounter difficulty applying due to lack of accessibility, language barriers, or inability to pay fees (in the present or the expected future). Additionally, there may not be a need to emphasize smaller class sizes or increased spending per student.

Appendix: Code for each question

Question 1

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from scipy import stats
from matplotlib import style
style.use('seaborn')

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

```
#%%
# 1: correlation btwn # applications and admissions
q1 = np.corrcoef(data['applications'], data['acceptances'])
# correlation = 0.802
q1_sq = q1[0][1]**2 # R^2 = 0.643

# plot
plt.scatter(data['applications'], data['acceptances'])
plt.title('Number of Applications vs. Acceptances to HSPHS - '
          + 'R^2 = {:.3f}'.format(q1_sq))
plt.xlabel('Applications to HSPHS')
plt.ylabel('Acceptances to HSPHS')
```

Question 2

```
#%%
# 2: better predictor - # of applications or application *rate*?
# raw # of applications:
q2a = q1[0][1]**2 # R^2 = 0.643

# clean data: missing school size data
data2 = data.copy()
data2 = data2[data2['school_size'].notnull()]
data2 = data2.reset_index(drop=True)

# application rate:
data2['application_rate'] = data2['applications']/data2['school_size']
data2['acceptance_rate'] = data2['acceptances']/data2['school_size']
rate_r = np.corrcoef(data2['application_rate'], data2['acceptance_rate'])
# correlation = 0.697
q2b = rate_r[0][1]**2 # R^2 = 0.485
# it appears that application # is a better predictor.

# plot rates
plt.scatter(data2['application_rate'], data2['acceptance_rate'])
plt.title('Rate of Applications vs. Acceptances to HSPHS - '
          + 'R^2 = {:.3f}'.format(q2b))
plt.xlabel('Rate of Applications to HSPHS')
plt.ylabel('Rate of Acceptances to HSPHS')
```

Question 3

```
#%%
# 3: which school has best *per student* odds of sending someone to HSPHS?
# odds = happening vs not happening ratio. p : ~p

data3 = data.copy()
data3['odds'] = data3['acceptances']/(data3['applications']
                                      -data3['acceptances'])
bestodds = data3['odds'].max()
q3 = data.loc[data3['odds']==bestodds, 'school_name']
# THE CHRISTA MCAULIFFE SCHOOL\I.S. 187 has the best odds

data3['psodds'] = data3['odds']/data3['acceptances']
bestpsodds = data3['psodds'].max()
q3a = data.loc[data3['psodds']==bestpsodds, 'school_name']
# SPECIAL MUSIC SCHOOL if normalized by # of applications

n = data3.shape[0]
data3 = data3.sort_values(by='psodds', ascending=False)

# plot:
plt.bar(np.linspace(1,n,n), data3['psodds'], width=0.9)
plt.title('Per Student Odds of Going to HSPHS')
plt.xlabel('Schools, ordered by decreasing per student odds')
plt.ylabel('Per Student Odds')
```

Question 4

```
#%%
# 4: is there a relationship btwn how students perceive their school (L-Q) and
# how school performs objectively (V-X)?

from sklearn.decomposition import PCA

# clean the rows of interest of the data
data4 = data.copy()
data4 = data4.loc[:,'rigorous_instruction':'trust'.join(data4.loc[:,'student_a

data4 = data4[data4.loc[:,'rigorous_instruction':'trust'].notnull().all(1)]
data4 = data4[data4.loc[:,'student_achievement':'math_scores_exceed'].notnull()
```

```
#%%
# 4a: PCA on first 6 factors

data4a = data4.loc[:,'rigorous_instruction':'trust']

zscored_data = stats.zscore(data4a)

pca = PCA()
pca.fit(zscored_data)
```

Question 4 cont.

```
eig_vals = pca.explained_variance_
loadings = pca.components_
rotated_data1 = pca.fit_transform(zscored_data)
covar_explained = eig_vals/sum(eig_vals)*100

num_classes = len(eig_vals)
plt.bar(np.linspace(1,num_classes,num_classes),eig_vals)
plt.xlabel('Principal component')
plt.ylabel('Eigenvalue')
plt.title('Scree Plot for School Climate Factors PCA')
plt.plot([0,num_classes],[1,1],color='red',linewidth=1) # Kaiser criterion line
```

```
#%%
# interpret factors: plot principal components
which_principal_component = 0
pc = which_principal_component+1
plt.bar(np.linspace(1,6,6),loadings[which_principal_component])
plt.xlabel('Columns L-Q')
plt.ylabel('Loading')
plt.title('Loadings for PC {}'.format(pc))
```

```
#%%
# 4b: PCA on last 3 factors

data4b = data4.loc[:,'student_achievement':'math_scores_exceed']
zscored_data = stats.zscore(data4b)
pca = PCA()
pca.fit(zscored_data)
eig_vals = pca.explained_variance_
loadings = pca.components_
rotated_data2 = pca.fit_transform(zscored_data)
covar_explained = eig_vals/sum(eig_vals)*100

num_classes = len(eig_vals)
plt.bar(np.linspace(1,num_classes,num_classes),eig_vals)
plt.xlabel('Principal component')
plt.ylabel('Eigenvalue')
plt.title('Scree Plot for Objective Measures of Achievement Factors PCA')
plt.plot([0,num_classes],[1,1],color='red',linewidth=1) # Kaiser criterion line
```

```
#%%
# interpret factors: plot principal components
which_principal_component = 0
pc = which_principal_component+1
plt.bar(np.linspace(1,3,3),loadings[which_principal_component])
plt.xlabel('Columns V-X')
plt.ylabel('Loading')
plt.title('Loadings for PC {}'.format(pc))
```

Question 4 cont.

```
#%%
# 4c = correlate the two PCs
q4 = np.corrcoef(rotated_data1[:,0], rotated_data2[:,0])
# r = -0.367
# plot:
plt.scatter(rotated_data1[:,0], rotated_data2[:,0])
plt.title('PC: School Climate vs. PC: Objective Measures of Achievement')
plt.xlabel('School Climate')
plt.ylabel('Objective Measures of Achievement')
# weak negative relationship
```

Question 5

```
#%%
# 5: do charter schools have higher average standardized test (ST) scores?
# hypothesis test! use the framework

# H0 = There is no significant difference between the average ST scores of
# charter and non-charter schools.
# Test: independent samples t-test

# clean, set up data
data5 = data.copy()
data5 = data5.loc[:,'dbn':'school_name'].join(data5.loc[:,'student_achievement
data5 = data5[data5['student_achievement'].notnull()]
data5 = data5.reset_index(drop=True)

regschool = data5[0:470]
charter = data5[470:]

meanreg = np.mean(regschool['student_achievement']) # 3.37
meanchart = np.mean(charter['student_achievement']) # 3.63
t1,p1 = stats.ttest_ind(regschool['student_achievement'],
                        charter['student_achievement'])
# t = -3.071
# p = 0.002 < 0.01
# so at the 0.01 significance level, we can reject the H0
# and conclude that charter schools have significantly different
# ST scores than other schools.

# plot bars:
regschool_mean = np.mean(regschool['student_achievement'])
regschool_std = np.std(regschool['student_achievement'])
charter_mean = np.mean(charter['student_achievement'])
charter_std = np.std(charter['student_achievement'])

plt.bar(['Non-charter Schools','Charter Schools'],[regschool_mean, charter_mea
#        yerr=[regschool_std, charter_std], capsize=10)
plt.title('Student Achievement at Non-charter vs. Charter Schools')
plt.xlabel('Type of School')
plt.ylabel('Average Standardized Test Score')
```

Question 6

```
#%%
# 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?
# correlate spending with achievement

# spending & class size vs average ST score
data6 = data.copy()
data6 = data6.loc[:,'dbn':'avg_class_size'].join(data6.loc[:,'student_achieveme
data6 = data6[data6['per_pupil_spending'].notnull()]
data6 = data6[data6['student_achievement'].notnull()]
data6 = data6[data6.dbn != '02M347']

q6a = np.corrcoef(data6['per_pupil_spending'],data6['student_achievement'])
# -0.158 --> -0.218 when outlier removed
q6b = np.corrcoef(data6['avg_class_size'],data6['student_achievement'])
# 0.209

# ... vs acceptance rate
data6a = data.copy()
data6a = data6a.loc[:,'dbn':'avg_class_size'].join(data6a.loc[:,'school_size']]
data6a = data6a[data6a['per_pupil_spending'].notnull()]
data6a['rate'] = data6a['acceptances']/data6a['school_size']
data6a['rate'] = data6a['rate'].fillna(0)
data6a = data6a[data6a.dbn != '02M347']

q6c = np.corrcoef(data6a['per_pupil_spending'],data6a['rate'])
# -0.307 without outlier
q6d = np.corrcoef(data6a['avg_class_size'],data6a['rate'])
# 0.348

# ... vs # of acceptances
q6e = np.corrcoef(data6a['per_pupil_spending'],data6a['acceptances'])
# -0.360 --> -0.356 without outlier
q6f = np.corrcoef(data6a['avg_class_size'],data6a['acceptances'])
# 0.350

# ... vs reading score
data6b = data.copy()
data6b = data6b[data6b['per_pupil_spending'].notnull()]
data6b = data6b[data6b['reading_scores_exceed'].notnull()]
data6b = data6b[data6b['math_scores_exceed'].notnull()]
data6b = data6b[data6b.dbn != '02M347']

q6g = np.corrcoef(data6b['per_pupil_spending'],data6b['reading_scores_exceed']]
# -0.498 --> -0.529 without outlier
q6h = np.corrcoef(data6b['avg_class_size'],data6b['reading_scores_exceed'])
# 0.537
```

Question 6 cont.

```
# ... vs math score
q6i = np.corrcoef(data6b['per_pupil_spending'],data6b['math_scores_exceed'])
# -0.485 --> -0.515 without outlier
q6j = np.corrcoef(data6b['avg_class_size'],data6b['math_scores_exceed'])
# 0.558
```

```
#%%
# plot
plt.scatter(data6b['avg_class_size'], data6b['math_scores_exceed'])
plt.title('Average Class Size vs. Math Scores')
plt.xlabel('Average Class Size')
plt.ylabel('Proportion of Math Scores Above State Standards')
```

```
#%%
plt.scatter(data6b['per_pupil_spending'], data6b['reading_scores_exceed'])
plt.title('Per Pupil Spending vs. Reading Scores')
plt.xlabel('Per Pupil Spending')
plt.ylabel('Proportion of Reading Scores Above State Standards')
```

```
#%%
# Do schools with larger classes have different results
# in math than schools with smaller classes?
# H0: No difference.

data6h = data.copy()
data6h = data6h.loc[:,'dbn':'avg_class_size'].join(data6h.loc[:,'math_scores_ex
data6h = data6h[data6h['avg_class_size'].notnull()]
data6h = data6h[data6h['math_scores_exceed'].notnull()]
data6h = data6h.reset_index(drop=True)

medsize = np.median(data6h['avg_class_size'])
data6h['larger_avg_class_size'] = 0
data6h['larger_avg_class_size'] = (data6h['avg_class_size']>=medsize).astype(in

larger = data6h.loc[data6h['larger_avg_class_size']==1]
smaller = data6h.loc[data6h['larger_avg_class_size']==0]

# acceptance rate
t5,p5 = stats.ttest_ind(larger['math_scores_exceed'], smaller['math_scores_exce
# t = 13.05, p = 0 (1.897892698546232e-33)
# The p-value is essentially zero. We can reject the null hypothesis.

# plot bars
larger_mean = np.mean(larger['math_scores_exceed'])
smaller_mean = np.mean(smaller['math_scores_exceed'])

plt.bar(['Larger (Above Median)','Smaller (Below Median)'],
        [larger_mean, smaller_mean])
plt.title('Math Scores Exceeding State Expectations Based on Average Class Size
plt.xlabel('Average Class Size')
plt.ylabel('Proportion of Math Scores Exceeding Standards')
```

Question 6 cont.

```
#%%
# Do schools with greater spending per student have
# different reading score achievements than schools that spend
# less per student?
# H0: No difference.

data6ha = data.copy()
data6ha = data6ha.loc[:,'dbn':'per_pupil_spending'].join(data6ha.loc[:,'reading
data6ha = data6ha[data6ha['per_pupil_spending'].notnull()]
data6ha = data6ha[data6ha['reading_scores_exceed'].notnull()]
data6ha = data6ha.reset_index(drop=True)

medspend = np.median(data6ha['per_pupil_spending'])
data6ha['rich'] = 0
data6ha['rich'] = (data6ha['per_pupil_spending']>=medspend).astype(int)

rich = data6ha.loc[data6ha['rich']==1]
poor = data6ha.loc[data6ha['rich']==0]

# acceptance rate
t6,p6 = stats.ttest_ind(rich['reading_scores_exceed'],poor['reading_scores_exce
# t = -12.31, p = 0 (2.1383053064464824e-30)
# The p-value is essentially zero. We can reject the null hypothesis.

# plot bars
rich_mean = np.mean(rich['reading_scores_exceed'])
poor_mean = np.mean(poor['reading_scores_exceed'])

plt.bar(['Richer (Above Median)','Poorer (Below Median)'],
        [rich_mean, poor_mean])
plt.title('Reading Scores Exceeding State Expectations Based on Per Pupil Spend
plt.xlabel('Per Pupil Spending')
plt.ylabel('Proportion of Reading Scores Exceeding Standards')
```

Question 7

```
#%%
# 7: what proportion of schools accounts for 90% of all students
# accepted to HSPHS?
data7 = data.copy()
data7 = data7.sort_values(by='acceptances', ascending=False)
data7 = data7.reset_index(drop=True)
total = sum(data7['acceptances'])
threshold = total*0.9
sum_accept = 0
for i in range(len(data7['acceptances'])):
    sum_accept += data7['acceptances'][i]
    if sum_accept > threshold:
        threshold_index = i
        break
# index = 122,
# so 123 schools account for 90% of all students accepted.
sum_90 = sum(data7['acceptances'][0:123])
proportion_90 = 123/594
# proportion = 0.207 --> 20.7% of schools
# account for 90% of all students accepted.
```

```
#%%
# bar graph for #7
# rank-ordered by decreasing # of acceptances to HSPHS.
n = data7.shape[0]
# plot:
plt.bar(np.linspace(1,n,n), data7['acceptances'], width=0.9)
plt.title('Schools, ordered by decreasing number of acceptances to HSPHS')
plt.xlim(-10,n)
plt.xlabel('School rank (most to least acceptances)')
plt.ylabel('Number of acceptances')
plt.axvline(x=threshold_index+1, label='90% of acceptances', color='r')
plt.legend()
```

Question 8

```
#%%
# 8: build a model including all factors as to what school characteristics
# are most important in sending students to HSPHS (acceptance # or rate)
# and achieving high scores on objective measures of achievement (V-W).

from sklearn.decomposition import PCA

# select rows
data8 = data.copy()
data8 = data8.dropna()
data8a = data8.loc[:,'per_pupil_spending':'avg_class_size']
data8b = data8.loc[:,'multiple_percent':'white_percent'].join(data8.loc[:,'asia
data8bb = data8.loc[:,'black_percent':'hispanic_percent']
data8c = data8.loc[:,'rigorous_instruction':'trust']
data8d = data8.loc[:,'disability_percent':'ESL_percent']
data8e = data8.loc[:,'student_achievement':'math_scores_exceed']
```

Question 8 cont.

```
#%%
# z-score data:
zscored_data = stats.zscore(data8e) #change data #

# run the PCA:
pca = PCA()
pca.fit(zscored_data)

# return outputs:
eig_vals = pca.explained_variance_
loadings = pca.components_
rotated_data6 = pca.fit_transform(zscored_data) # change # after rotated_data
covar_explained = eig_vals/sum(eig_vals)*100
# scree plot:
num_classes = len(eig_vals)
plt.bar(np.linspace(1,num_classes,num_classes),eig_vals)
plt.xlabel('Principal component')
plt.ylabel('Eigenvalue')
plt.plot([0,num_classes],[1,1],color='red',linewidth=1)
```

```
#%%
# interpret factors: plot principal components
which_principal_component = 0
pc = which_principal_component+1
plt.bar(np.linspace(1,3,3),loadings[which_principal_component]) # change #s
plt.xlabel('Object Achievement Factors') # change to combined variable name
plt.ylabel('Loading')
plt.title('Loadings for PC {}'.format(pc))
```

```
#%%
# put all combined factors in one df
df = pd.DataFrame({'applications': data8['applications'],
                   'acceptances': data8['acceptances'],
                   'per_pupil_spending': data8['per_pupil_spending'],
                   'avg_class_size': data8['avg_class_size'],
                   'material_resources': rotated_data2[:,0],
                   'diversity_amw': rotated_data3a[:,0],
                   'diversity_bh': rotated_data3b[:,0],
                   'school_climate': rotated_data4[:,0],
                   'disadvantages': rotated_data5[:,0],
                   'school_size': data8['school_size'],
                   'objective_achievement': rotated_data6[:,0]})
```

```
#%%
# multiple regression #1: number of acceptances
style.use('seaborn')
# Descriptives:
D1 = np.mean(df,axis=0) # take mean of each column
D2 = np.median(df,axis=0) # take median of each column
D3 = np.std(df,axis=0) # take std of each column
```

Question 8 cont.

```
# Model: All factors
from sklearn import linear_model
X = np.transpose([df['per_pupil_spending'], df['avg_class_size'],
                  df['diversity_amw'], df['diversity_bh'],
                  df['school_climate'], df['disadvantages'],
                  df['school_size']]) # predictors
Y = df['acceptances'] #
regr1 = linear_model.LinearRegression()
regr1.fit(X,Y) # use fit method
r_sqr1 = regr1.score(X,Y) # R^2
betas1 = regr1.coef_ # m
y_int1 = regr1.intercept_  # b

# Visualize:
y_hat1 = (betas1[0]*df['per_pupil_spending'] + betas1[1]*df['avg_class_size']
          + betas1[2]*df['diversity_amw'] + betas1[3]*df['diversity_bh']
          + betas1[4]*df['school_climate'] + betas1[5]* df['disadvantages']
          + betas1[6]*df['school_size'] + y_int1)
plt.plot(y_hat1,df['acceptances'],'o',markersize=7) # acceptances
plt.xlabel('Prediction from model')
plt.ylabel('Actual number of acceptances')
plt.title('Predicting Number of Acceptances - R^2: {:.3f}'.format(r_sqr1))
```

```
#%%
# multiple regression #2: objective measures of achievement

# Model
from sklearn import linear_model
X = np.transpose([df['per_pupil_spending'], df['avg_class_size'],
                  df['diversity_amw'], df['diversity_bh'],
                  df['school_climate'], df['disadvantages'],
                  df['school_size']]) # predictors
Y = df['objective_achievement'] #
regr2 = linear_model.LinearRegression()
regr2.fit(X,Y) # use fit method
r_sqr2 = regr2.score(X,Y) # R^2
betas2 = regr2.coef_ # m
y_int2 = regr2.intercept_  # b

# Visualize:
y_hat2 = (betas2[0]*df['per_pupil_spending'] + betas2[1]*df['avg_class_size']
          + betas2[2]*df['diversity_amw'] + betas2[3]*df['diversity_bh']
          + betas2[4]*df['school_climate'] + betas2[5]* df['disadvantages']
          + betas2[6]*df['school_size'] + y_int2)
plt.plot(y_hat2,df['objective_achievement'],'o',markersize=7) # achievement
plt.xlabel('Prediction from model')
plt.ylabel('Actual achievement')
plt.title('Predicting Objective Achievement - R^2: {:.3f}'.format(r_sqr2))
```

Question 8 cont.

```
#%%
# multiple regression #3: number of applications

# Model
from sklearn import linear_model
X = np.transpose([df['per_pupil_spending'], df['avg_class_size'],
                  df['diversity_amw'], df['diversity_bh'],
                  df['school_climate'], df['disadvantages'],
                  df['school_size']]) # predictors
Y = df['applications'] #
regr3 = linear_model.LinearRegression()
regr3.fit(X,Y) # use fit method
r_sqr3 = regr3.score(X,Y) # R^2
betas3 = regr3.coef_ # m
y_int3 = regr3.intercept_  # b

# Visualize:
y_hat3 = (betas3[0]*df['per_pupil_spending'] + betas3[1]*df['avg_class_size']
          + betas3[2]*df['diversity_amw'] + betas3[3]*df['diversity_bh']
          + betas3[4]*df['school_climate'] + betas3[5]* df['disadvantages']
          + betas3[6]*df['school_size'] + y_int3)
plt.plot(y_hat3,df['applications'],'o',markersize=7) # y_hat, applications
plt.xlabel('Prediction from model')
plt.ylabel('Actual number of applications')
plt.title('Predicting Number of Applications - R^2: {:.3f}'.format(r_sqr3))
```