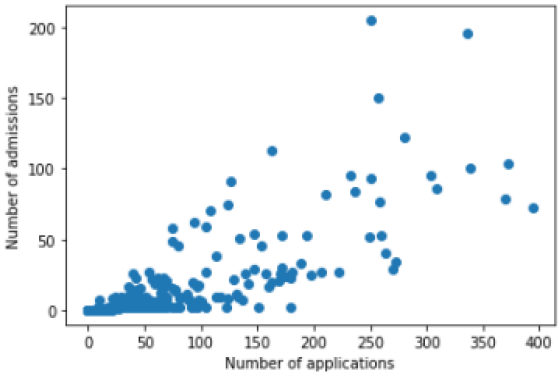
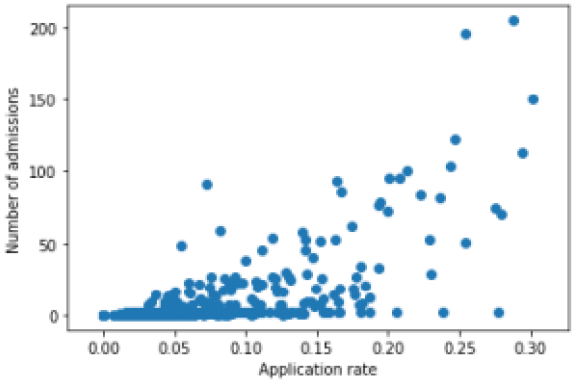
Oviya Adhan
Dr. Pascal Wallisch
Introduction to Data Science
22 December 2020

IDS Final Project

1) In order to find the correlation between the number of applications and admissions I turned each appropriate column into its own series and used corr() to find the correlations between the two variables. The correlation between the number of applications and admissions to HSPHS is r = 0.8017265370719315. This means that there is a strong correlation between the number of applications and the number of admissions. As more applications are sent from a school, more applications are accepted at HSPHS from that school.



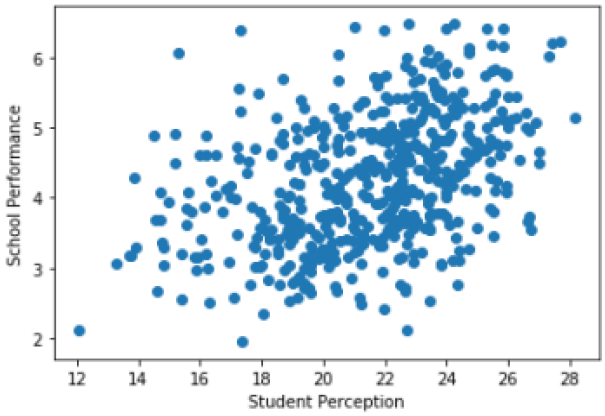
2) In order to compare applications and application rates' correlation with acceptances, I did the same steps as Question 1, but with the application rate, which was calculated by dividing applications by the school size. The correlation between the application rate and admissions is r = 0.658750752900268. Being lower than 0.80171, the application rate has a lower correlation with admissions than the raw number of applications. Therefore, the raw number of applications from a school is a better predictor for admissions than the application rate.



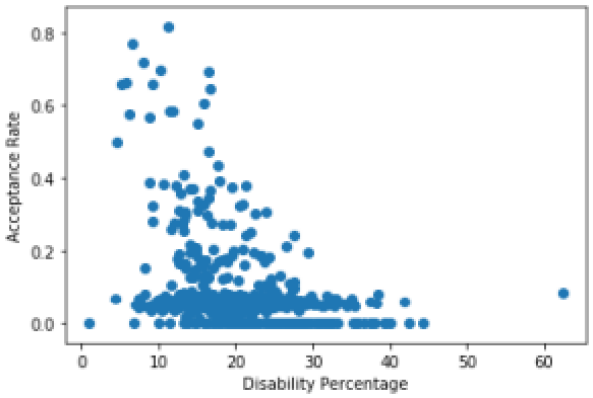
3) In order to find the best *per student* odds of sending someone to HSPHS, I first created a separate column with the per-student probability of getting into HSPHS by taking the acceptance column value and dividing it by the school size column. Then, using that column, I used the max() function to find that the highest per-students probability of getting accepted from a single school is 0.23482245131729668. Using the highest per-student odds, I located the index where the per

student odds equaled the max_value, and used that index to locate the school name using loc(). The result is The Christa McCauliffe School.

4) Finding the relationship between student perceptions about their schools and school performance based on achievements required that there were separate columns created in the dataframe combining the different factors that make up each category. For student perception, the columns that represent rigorous instruction, collaborative teachers, supportive environment, effective school leadership, and strong family community were added together to create singular student perceptions values. For school performance, the columns representing students achievement, reading scores exceed, and math scores exceed were added together to create singular school performance values. Then the correlation between the two new values was calculated, which turned out to be r = 0.41874194039783613.

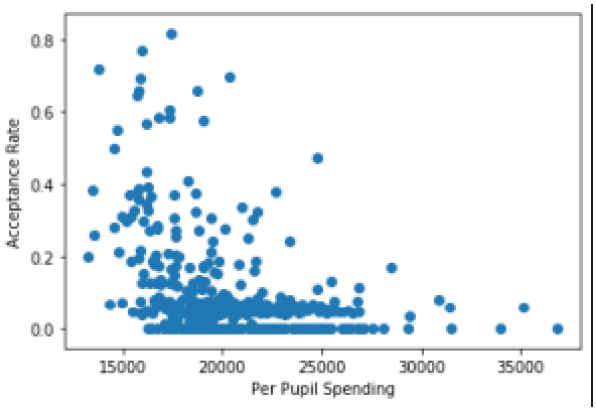


5) While looking through the variables available to compare, the disability percentage stood out to me. Disability can include learning disabilities as well, which would logically result in lower test scores. I wanted to test whether a higher disability percentage would lead to a lower acceptance rate, which may not be as strong as schools with lower disability percentages. The hypothesis that I decided to test was the higher the percentage rate of students with disabilities (independent variable), the lower the acceptance rate to HSPHS would be (dependent variable). In order to do this, I simply calculated the acceptance rate by dividing the acceptances by the number of applications and found the correlation between that value and disability percentage, which turned out to be r = -0.37334474181581684, which shows a moderate negative correlation between the two variables, meaning that my hypothesis that the higher the disability percentage, the lower the acceptance rate to HSPHS is possibly correct.



6) In order to see whether the availability of material resources impacts objective measures of achievement or admission to HSPHS, I chose to find the correlation between acceptance rate

and per pupil spending. This yields a correlation coefficient of r = -0.39581588119318084, which is a moderate correlation, suggesting that in some way, per pupil spending affects acceptance rate to HSPHS.



7) To find the proportion of schools that make up 90% of admissions to HSPHS, I first sorted the data from highest to lowest number of admissions to HSPHS. From there, I found the total number of admitted students to HSPHS, and found the threshold number that would be 90% of admitted students. I then iterated through the data, adding each admitted value onto a sum and adding 1 to the number of schools with a flag at the threshold. Once it hits the threshold the loop breaks and produces the number of schools it took to reach that threshold amount. With that number, I found the proportion by dividing it by the total number of schools. 0.765993265993266 or 76.5993265993266% of NYC middle schools make up 90% of the admissions to HSPHS.
8) In order to do a multiple regression, I decided to use a numpy array and matrices. Although my code did not yield any valid results, the steps I took where to find the linear regression of each possible dependent variable to each of the two independent variables, acceptances and school performance (a combination of student achievement, reading scores, and math scores), then did a regression model with all the factors combined. The code I attempted is below in the appendix.
9) The school characteristics that appear most relevant in acceptances to HSPHS are applications and application rate and student perception of their own school. This is due to the higher correlation between these variables and the acceptances to HSPHS. I expected to find a positive correlation between acceptance rate and per pupil spending, but instead there is a moderate negative correlation, suggesting that as per pupil spending goes up, acceptance rate goes down. Perhaps, more spending means a richer and better performing school district, so even the high-achievers at these schools would choose to continue their education with their own highly-funded public high school instead of HSPHS.
10) I would recommend three things to the New York City Department of Education:
    a) Spread awareness about HSPHS to get application numbers to be higher. The largest correlation is between applications and acceptances, so if there are simply more applications from a school, there should be higher acceptances.
    b) Since the correlation between student perception of their school and school performance is pretty moderate, I would suggest focusing on school pride for each of these schools. Increasing community feel especially would help students feel more comfortable in their learning environment and would result in better performance.
    c) Finally, in order to boost both acceptances and school performance, I would suggest hiring more teachers in schools with large class sizes to lower that number. That way, students would receive more 1-on-1 attention, and school performance overall should increase.
11)

Appendix (Code)

Question 1:

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Mon Dec 21 19:19:22 2020

@author: oviya
"""

import numpy as np
import pandas as pd

import matplotlib.pyplot as plt


df = pd.read_csv("middleSchoolData.csv")
data = np.genfromtxt(df, delimiter = ',')


# Q1: Correlation b/w applications and admissions

applications = df["applications"]
admissions = df["acceptances"]
app_ad_corr = applications.corr(admissions)
corr = str(app_ad_corr)

print("Question 1: The correlation between the number of applications and admissions is " + corr)

# Plot
f1 = plt.figure(1)
x = df.applications
y = df.acceptances
plt.scatter(x,y)
plt.xlabel('Number of applications')
plt.ylabel('Number of admissions')


```

Question 2:

```

# Q2: Which is better predictor for admission to HSPS: raw number of applications or application rate

df['application_rate'] = df['applications']/df['school_size']
application_rate = df['application_rate']
apprate_ad_corr = application_rate.corr(admissions)
corr2 = str(apprate_ad_corr)
print("\nQuestion 2: The correlation between the application rate and admissions is " + corr2)

## Compare correlations
if app_ad_corr > apprate_ad_corr:
    print("Raw number of applications is a better predictor for admissions.")
else:
    print ("Application rate is a better predictor for admissions.")


# Plot
f2 = plt.figure(2)
a = df.application_rate
b = df.acceptances
plt.scatter(a,b)
plt.xlabel('Application rate')
plt.ylabel('Number of admissions')


```

Question 3:

```
# Q3: Best per-student odds of sending someone to HSPHS

df['per_student_odds'] = df['acceptances']/df['school_size']
per_student_odds = df['per_student_odds']
max_value = per_student_odds.max()
max_val_str = str(max_value)
print("\nQuestion 3: The highest per-student odds of sending someone to HSPHS is " + max_val_str)


name_odds = df[['school_name', 'per_student_odds']]
index = df[df['per_student_odds'] == max_value].index.values
name = df.loc[index, 'school_name']
print(name)
print("The school with the highest per-student odds of sending a student to HSPHS is shown above.")
```

Question 4:

```
# Q4: Relationship b/w students' perception of school and school performance

df['student_perception'] = df['rigorous_instruction'] + df['collaborative_teachers'] + \
df['supportive_environment'] + df['effective_school_leadership'] + df['strong_family_community_ties'] + df['trust']
student_perception = df['student_perception']
df['school_performance'] = df['student_achievement'] + df['reading_scores_exceed'] + df['math_scores_exceed']
school_performance = df['school_performance']
perc_perf_corr = student_perception.corr(school_performance)
corr3 = str(perc_perf_corr)
print("\nQuestion 4: The correlation between student perception and school performance is " + corr3)

abs_corr = abs(perc_perf_corr)

if abs_corr == 1:
    print("There is a perfect correlation between student perception and school performance.")
elif abs_corr >= 0.5:
    print("There is a high/strong correlation between student perception and school performance.")
elif abs_corr >= 0.3:
    print("There is a moderate correlation between student perception and school performance.")
elif abs_corr >= 0:
    print("There is a low/weak correlation between student perception and school performance.")
else:
    print("Error: Correlation is out of range.")


# Plot
f3 = plt.figure(3)
j = df.student_perception
k = df.school_performance
plt.scatter(j,k)
plt.xlabel('Student Perception')
plt.ylabel('School Performance')
```

Question 5:

```
# Q5: Testing a custom hypothesis: The higher the disability percent, the lower the acceptance rate

df['acceptance_rate'] = df['acceptances']/df['applications']
acceptance_rate = df['acceptance_rate']
disability = df['disability_percent']
accrate_dis_corr = acceptance_rate.corr(disability)
corr4 = str(accrate_dis_corr)
print("\nQuestion 5: The correlation between acceptance rate and disability percentage is " + corr4)

abs_corr4 = abs(accrate_dis_corr)

if abs_corr4 == 1:
    print("There is a perfect correlation between acceptance rate and disability percentage.")
elif abs_corr4 >= 0.5:
    print("There is a high/strong correlation between acceptance rate and disability percentage.")
elif abs_corr4 >= 0.3:
    print("There is a moderate correlation between acceptance rate and disability percentage.")
elif abs_corr4 >= 0:
    print("There is a low/weak correlation between acceptance rate and disability percentage.")
else:
    print("Error: Correlation is out of range.")

# Plot
f4 = plt.figure(4)
q = df.disability_percent
r = df.acceptance_rate
plt.scatter(q,r)
plt.xlabel('Disability Percentage')
plt.ylabel('Acceptance Rate')
```

Question 6:

```
# Q6: Finding evidence that the availibility of resources (per student spending) impacts achievement (admissions in this case)

df['acceptance_rate'] = df['acceptances']/df['applications']
acceptance_rate = df['acceptance_rate']
spending = df['per_pupil_spending']
spending_accrate_corr = acceptance_rate.corr(spending)
corr5 = str(spending_accrate_corr)
print("\nQuestion 6: The correlation between acceptance rate and per-student spending is " + corr5)

abs_corr5 = abs(spending_accrate_corr)

if abs_corr5 == 1:
    print("There is a perfect correlation between acceptance rate and disability percentage.")
elif abs_corr5 >= 0.5:
    print("There is a high/strong correlation between acceptance rate and disability percentage.")
elif abs_corr5 >= 0.3:
    print("There is a moderate correlation between acceptance rate and disability percentage.")
elif abs_corr5 >= 0:
    print("There is a low/weak correlation between acceptance rate and disability percentage.")
else:
    print("Error: Correlation is out of range.")

# Plot
f5 = plt.figure(5)
a = df.per_pupil_spending
b = df.acceptance_rate
plt.scatter(a,b)
plt.xlabel('Per Pupil Spending')
plt.ylabel('Acceptance Rate')

```

Question 7:

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

# Sorted acceptances list in descending order
accept = df.sort_values('acceptances', ascending=False)['acceptances']
# Get number of acceptances that make up 90%
total_acceptances = df['acceptances']
threshold = total_acceptances.sum() * 0.9
total_schools = len(accept)
num_schools = 0
num_acceptances = 0

for i in range(len(accept)):
    if num_acceptances < threshold:
        num_acceptances += accept[i]
        num_schools += 1
    else:
        break
    prop = num_schools/ total_schools
    perc = prop * 100

prop_str = str(prop)
perc_str = str(perc)

print("\nQuestion 7: " + prop_str + " or  " + perc_str + "% of all NYC schools make up 90% of admissions to HSPHS.")


```

Question 8:

```
#%% Q8: Multiple regression model

from sklearn import linear_model

X_df = df[['applications', 'acceptances', 'per_pupil_spending', 'avg_class_size', 'asian_percent', 'black_percent', \
        'hispanic_percent', 'multiple_percent', 'white_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']]
Y1_df = df[['acceptances']]
Y2_df = df[['school_performance']]


# Desriptives
d1 = np.mean(data,axis=0)
d2 = np.median(data,axis=0)
d3 = np.std(data,axis=0)
d4 = np.corrcoef(data[:,0],data[:,1])


# For acceptances regression
X1 = np.transpose([data[:,2],data[:,4],data[:,5], data[:,6], data[:,6], data[:,7], data[:,8], data[:,9], \
                data[:,10], data[:,11], data[:,12], data[:,13], data[:,14], data[:,15], data[:,16], \
                data[:,17], data[:,18], data[:,19], data[:20], data[:,21], data[:,22], data[:,23]])
Y1 = data[:,3]
regr = linear_model.LinearRegression()
regr.fit(X1,Y1)
r_sqr = regr.score(X1,Y1)
betas = regr.coef_
y_int = regr.intercept_
# Plot
y_hat = betas[0]*data[:,0] + betas[1]*data[:,1] + betas[2]*data[:,2] + y_int
plt.plot(y_hat,data[:,3],'o',markersize=.75)
plt.xlabel('Prediction')
plt.ylabel('Actual acceptances')
plt.title('R^2: {:.3f}'.format(r_sqr))


# For achievement regression
X2 = np.transpose([data[:,2], data[:,3], data[:,4],data[:,5], data[:,6], data[:,6], data[:,7], data[:,8], \
                data[:,9], data[:,10], data[:,11], data[:,12], data[:,13], data[:,14], data[:,15], data[:,16], \
                data[:,17], data[:,18], data[:,19], data[:20]])
Y2 = np.transpose([data[:,21], data[:,22], data[:,23]])
regr = linear_model.LinearRegression()
regr.fit(X2,Y2)
r_sqr = regr.score(X2,Y2)
betas = regr.coef_
y_int = regr.intercept_
# Plot
y_hat = betas[0]*data[:,0] + betas[1]*data[:,1] + betas[2]*data[:,2] + y_int
plt.plot(y_hat,data[:,21],'o',markersize=.75)
plt.xlabel('Prediction')
plt.ylabel('Actual school performance')
plt.title('R^2: {:.3f}'.format(r_sqr))

```

Output:

```
In [155]: runfile('/Users/oviya/Desktop/Data Science/OviyaAdhan-FinalProject.py', wdir='/Users/oviya/Desktop/Data
Science')
Question 1: The correlation between the number of applications and admissions is 0.8017265370719315

Question 2: The correlation between the application rate and admissions is 0.658750752900268
Raw number of applications is a better predictor for admissions.

Question 3: The highest per-student odds of sending someone to HSPHS is 0.23482245131729668
304    THE CHRISTA MCAULIFFE SCHOOL\I.S. 187
Name: school_name, dtype: object
The school with the highest per-student odds of sending a student to HSPHS is shown above.

Question 4: The correlation between student perception and school performance is 0.41874194039783613
There is a moderate correlation between student perception and school performance.

Question 5: The correlation between acceptance rate and disability percentage is -0.37334474181581684
There is a moderate correlation between acceptance rate and disability percentage.

Question 6: The correlation between acceptance rate and per-student spending is -0.39581588119318084
There is a moderate correlation between acceptance rate and disability percentage.

Question 7: 0.765993265993266 or  76.5993265993266% of all NYC schools make up 90% of admissions to HSPHS.
```