Rodrigo Quijano-luna

5/17/2021

DS-UA 112: Introduction to Data Science

Prof. Pascal Wallisch

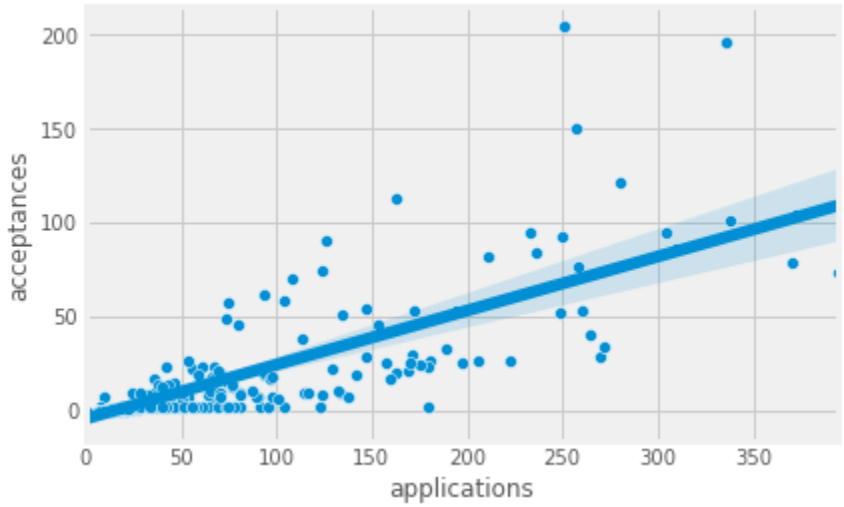
Data Project on New York City High Schools

Cleaning data involved the dropping of rows with null values:

```
df = df.dropna().reset_index().drop(columns=[ 'index'])
```

Principal Component Analysis followed a strict guideline, where I grouped the columns I wanted to be made into one vector and made that the component accounting for those variables. This effectively reduces the total number of columns from 24 columns to 17.

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



Output:

R-Squared: 0.6484718859141051

Correlation between applications and acceptance numbers: 0.805278

There is a clear correlation between the number of applications and number of acceptances into HSPHS. However, I am inclined to disregard the similarities between these two columns given that many schools have very few accepted students, let alone acceptances, meaning that the data is skewed and although the first few schools in the scatterplot are in fact correlated with the y axis value, the further you go, the less reliable this linear regression becomes.

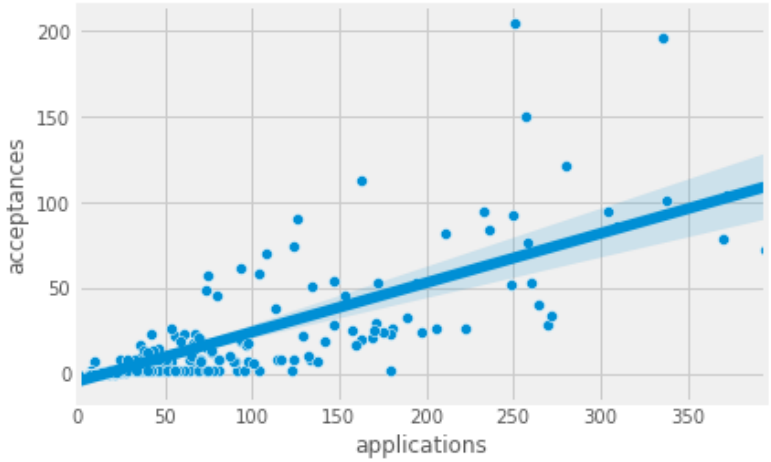
2) What is a better predictor of admission to HSPHS? Raw number of applications or application *rate*?

Acceptances v. Applications

R-Squared:

0.6484718859141051

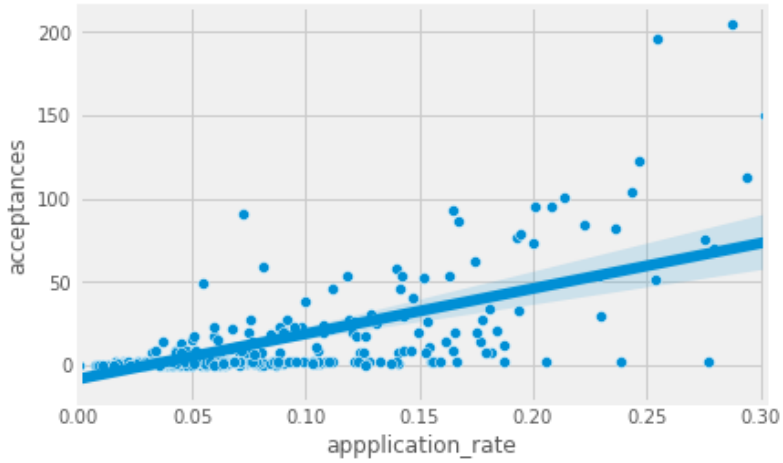
Correlation: 0.805278



Acceptances v. Application Rate

R-Squared: 0.4713576511390133

Correlation: 0.774239



Number of applications still has a better correlation to acceptance numbers than the application rate per school. This can probably mean that the acceptance rate for these HSPHS is more reliant on sheet number of applications rather than application rates.

R-Squared: 0.4715576511598133

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

Create a new column with acceptance rate, not by students who applied, but by every student in the school. (# of accepted / # of students in school)

```
472]: df['acceptance_per_student'] = df['acceptances']/df['school_size']
      display(df.sort_values(by='acceptance_per_student', ascending=False)[['school_name', 'acceptance_per_student']].head())
```

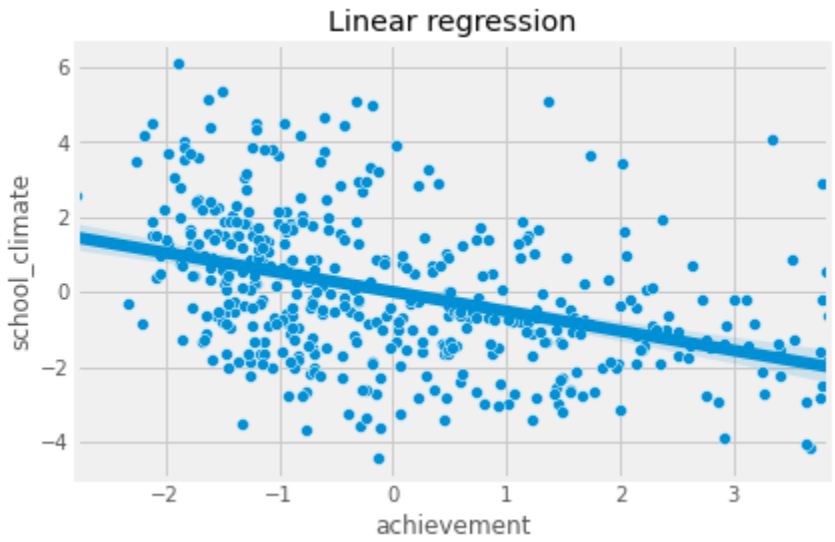
	school_name	acceptance_per_student
274	THE CHRISTA MCAULIFFE SCHOOL\I.S. 187	0.234822
20	NEW YORK CITY LAB MIDDLE SCHOOL FOR COLLABORAT...	0.203971
16	M.S. 255 SALK SCHOOL OF SCIENCE	0.181347
28	J.H.S. 054 BOOKER T. WASHINGTON	0.176056
10	EAST SIDE MIDDLE SCHOOL	0.166667

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

Acceptance rates per student are very high for these 5 schools, the best per student odds are for **The Christa Mcauliffe School.**

My method for finding acceptance odds per student was by dividing the number of acceptances per school by the school size entirely and this shows that Christa Mcauliffe School has a huge 23% acceptance rate into HSPHS.

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)?



We created two different vectors in the beginning using Principal Component

Analysis, creating a vector called school_climate and another for achievement.

R-Squared: 0.16020399105898336

It may seem that there is no relationship between these two vectors, but multiple regression in later models show that there is an important relationship between them two. School_climate actually is statistically significant in various other dependent variables.

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).

I am hypothesizing that schools below poverty median perform better than ones above the poverty median when it comes to admissions and achievement.

Null Hypothesis = Rich and poor schools perform the same.

Alternate Hypothesis = Rich schools out-perform poorer ones in both admissions to HSPHS and achievement scores.

T-Test to compare achievement scores for schools over and under median poverty_percent:

Ttest_indResult(statistic=-14.559292296328156, **p value=1.4048821653500924e-39**)

T-Test to compare acceptances per student for schools over and under median poverty_percent:

Ttest_indResult(statistic=-6.975522259318818, **p value=1.1031370079181595e-11**)

T-Test to compare school climate for schools over and under median poverty_percent:

Ttest_indResult(statistic=4.166330950246831, **p value=3.716823943881492e-05**)

We can see that using T-test for the means of two independent samples, there are significant differences in each of the three dependent variables I measured.

Printing out means for achievement of schools under and over the median of school_poverty:

Over Median	Under Median
-0.8061398661443846	**0.9012035296048078**

Those schools under the poverty_median percentage (AKA wealthier schools) have significantly better achievement metrics than those of schools over the median poverty percentage.

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?

I am hypothesizing that there is a difference between how schools manage resources per student and how those numbers affect achievement and admission scores to HSPHS.

Null Hypothesis = Schools spending more on students perform equally as well as those with less resources per student

Alternate Hypothesis = There is a difference between schools that spend more on students when compared to schools spending less per student than the median.

T-Test to compare achievement scores for schools over and under median spending per pupil:
Ttest_indResult(statistic=-10.765331107735816, **p value=3.4652961318758566e-24**)

T-Test to compare acceptances per student for schools over and under median spending per pupil:
Ttest_indResult(statistic=-6.653187464301445, **p value=8.405469631603547e-11**)

T-Test to compare school climate for schools over and under median spending per pupil:
Ttest_indResult(statistic=1.6209979417377685, **p value=0.10572323108313451**)

The results line up with my hypothesis that there are statistically significant differences between schools allotting less per student when compared to those schools that spend more per student. However, we can see that school climate isn't as different when comparing schools over and under the per_pupil_spending median.

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

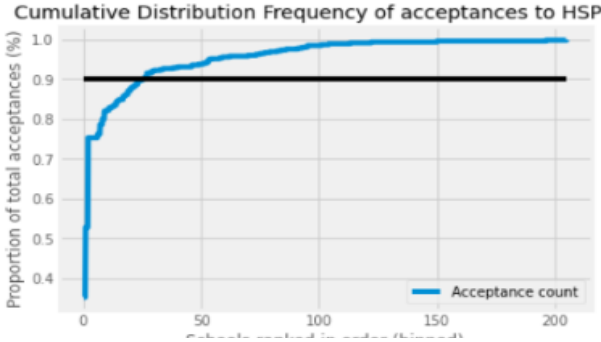
We can clearly see that very few schools account for the vast majority of acceptances into HSPHS, specifically, we can count that **20.27% of schools, which is about 91 of the schools** provided in the dataframe after having cleaned the dataset from NaN's.

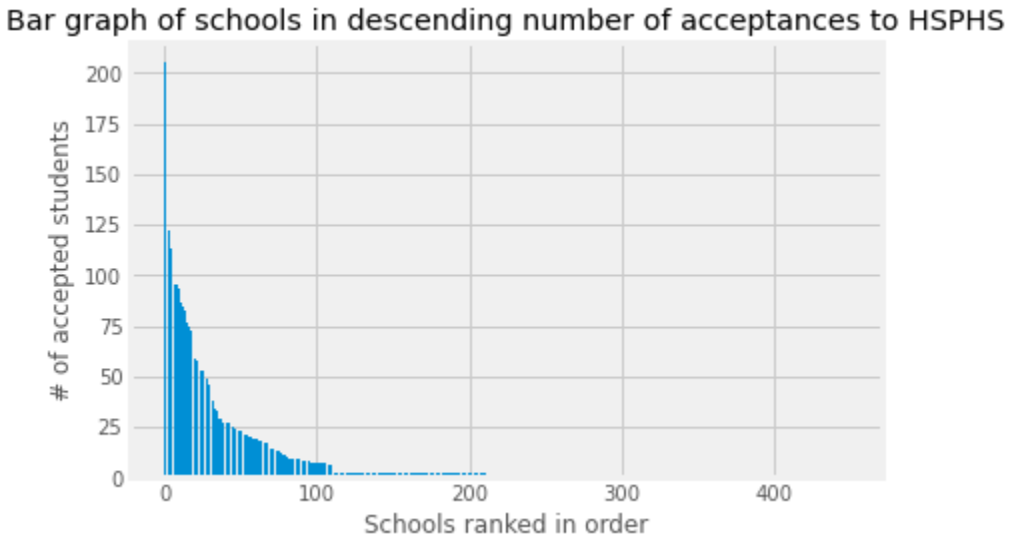
Let's see a bar graph showing the skewed distribution of the acceptance numbers for schools in descending order:

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

```
5]: df.sort_values(by='acceptances', ascending=False)
total = 0.9*df['acceptances'].sum()
school_count = 0
sum_acceptances = 0
while (sum_acceptances <= total):
    for index, row in df.sort_values(by='acceptances', ascending=False).iterrows():
        sum_acceptances += df.iloc[index]['acceptances']
        if sum_acceptances >= total:
            break
        else:
            school_count += 1
print(100*(round(school_count/len(df), 4)), '% of schools account for at least 90% of students accepted into HSPHS')
data = df['acceptances'].sort_values(ascending=False).values
count, bins_count = np.histogram(data, bins=1000)
pdf = count / sum(count)
cdf = np.cumsum(pdf)
plt.plot(bins_count[1:], cdf, label="Acceptance count")
plt.plot([0,205], [.9,.9], color='black')
plt.legend()
plt.gca().set_ylabel('Proportion of total acceptances (%)')
plt.gca().set_xlabel('Schools ranked in order (binned)')
plt.gca().set_title('Cumulative Distribution Frequency of acceptances to HSPHS ')
plt.show()
```

```
20.27 % of schools account for at least 90% of students accepted into HSPHS
```





The vast majority of schools don’t even have acceptances 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?

I've decided to begin with a multiple regression model to figure out what the significance of certain columns were, and these are my results for achievement score predictions.

[461]:

OLS Regression Results

Dep. Variable:	achievement	**R-squared:**	0.766
Model:	OLS	**Adj. R-squared:**	0.756
Method:	Least Squares	**F-statistic:**	78.03
Date:	Mon, 17 May 2021	**Prob (F-statistic):**	5.51e-123
Time:	12:26:48	**Log-Likelihood:**	-494.31
No. Observations:	449	**AIC:**	1027.
Df Residuals:	430	**BIC:**	1105.
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.9062	53.744	0.184	0.854	-95.728	115.540
applications	-0.0031	0.002	-1.386	0.167	-0.007	0.001
acceptances	0.0070	0.006	1.196	0.233	-0.005	0.019
per_pupil_spending	-2.506e-05	1.38e-05	-1.812	0.071	-5.22e-05	2.12e-06
avg_class_size	0.0155	0.010	1.620	0.106	-0.003	0.034
asian_percent	-0.0696	0.537	-0.130	0.897	-1.126	0.987
black_percent	-0.0987	0.537	-0.184	0.854	-1.154	0.957
hispanic_percent	-0.0894	0.537	-0.167	0.868	-1.145	0.966
multiple_percent	-0.0677	0.537	-0.126	0.900	-1.123	0.988
white_percent	-0.0829	0.537	-0.154	0.877	-1.138	0.973
rigorous_instruction	0.2647	0.091	2.919	0.004	0.086	0.443
poverty_percent	-0.0237	0.005	-5.018	0.000	-0.033	-0.014
ESL_percent	-0.0149	0.004	-3.546	0.000	-0.023	-0.007
school_size	0.0002	0.000	0.712	0.477	-0.000	0.001
school_climate	-0.0960	0.030	-3.252	0.001	-0.154	-0.038
appplication_rate	3.4285	1.646	2.083	0.038	0.193	6.664
acceptance_per_student	-1.6706	4.641	-0.360	0.719	-10.792	7.450
poverty_percent_median	0.0418	0.113	0.372	0.710	-0.179	0.263
over_spending_median	-0.1235	0.110	-1.124	0.262	-0.339	0.092

Omnibus:	114.046	**Durbin-Watson:**	1.998
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	415.118
Skew:	1.110	**Prob(JB):**	7.22e-91
Kurtosis:	7.154	**Cond. No.**	3.23e+07

We can see that there are 5 significant vectors (P-Value below 0.05) attributing to achievement scores. These are **rigorous_instruction, poverty_percent, ESL_percent, school_climate, and application_rate**. Application rate can be dismissed given that it's a column created by dividing two other columns.

Now, in terms of acceptance numbers:

OLS Regression Results

Dep. Variable:	acceptances	**R-squared:**	0.938
Model:	OLS	**Adj. R-squared:**	0.936
Method:	Least Squares	**F-statistic:**	384.7
Date:	Mon, 17 May 2021	**Prob (F-statistic):**	5.05e-248
Time:	12:26:49	**Log-Likelihood:**	-1438.2
No. Observations:	449	**AIC:**	2912.
Df Residuals:	431	**BIC:**	2986.
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	681.0039	438.153	1.554	0.121	-180.179	1542.187
applications	0.2649	0.013	20.621	0.000	0.240	0.290
per_pupil_spending	-7.268e-05	0.000	-0.643	0.520	-0.000	0.000
avg_class_size	0.0832	0.078	1.063	0.288	-0.071	0.237
asian_percent	-6.7968	4.381	-1.552	0.122	-15.407	1.813
black_percent	-6.7998	4.378	-1.553	0.121	-15.405	1.805
hispanic_percent	-6.8205	4.378	-1.558	0.120	-15.426	1.785
multiple_percent	-6.8644	4.378	-1.568	0.118	-15.468	1.740
white_percent	-6.7881	4.378	-1.550	0.122	-15.394	1.817
rigorous_instruction	-0.3815	0.741	-0.515	0.607	-1.838	1.075
poverty_percent	0.0703	0.038	1.830	0.068	-0.005	0.146
ESL_percent	-0.0459	0.034	-1.344	0.180	-0.113	0.021
school_size	-0.0080	0.002	-4.731	0.000	-0.011	-0.005
school_climate	-0.2558	0.241	-1.061	0.289	-0.730	0.218
appplication_rate	-157.3310	11.123	-14.145	0.000	-179.193	-135.469
acceptance_per_student	693.7745	17.960	38.629	0.000	658.474	729.075
poverty_percent_median	-0.7177	0.919	-0.781	0.435	-2.524	1.089
over_spending_median	1.9184	0.893	2.148	0.032	0.163	3.674

Omnibus:	233.012	**Durbin-Watson:**	2.126
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	11129.982
Skew:	1.473	**Prob(JB):**	0.00
Kurtosis:	27.212	**Cond. No.**	3.22e+07

We can see that in terms of acceptances into HSPHS, there are 5 significant factors, these include **applications, school_size, application_rate,acceptance_per_student, and over_spending_median**. Of course, application_rate and acceptances_per_student don't really matter in this case because their products of other columns. Let's see the results of doing OLS using the significant factors other than the two previously mentioned.

I also decided to add a different dependent variable into the previous two multiple regressions models, now trying to predict acceptance_per_student:

]:

OLS Regression Results

Dep. Variable:	acceptance_per_student	**R-squared:**	0.919
Model:	OLS	**Adj. R-squared:**	0.916
Method:	Least Squares	**F-statistic:**	286.8
Date:	Mon, 17 May 2021	**Prob (F-statistic):**	1.49e-222
Time:	12:35:13	**Log-Likelihood:**	1556.2
No. Observations:	449	**AIC:**	-3076.
Df Residuals:	431	**BIC:**	-3002.
Df Model:	17		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975
const	-1.0569	0.556	-1.902	0.058	-2.149	0.03
applications	-0.0003	1.91e-05	-13.802	0.000	-0.000	-0.00
acceptances	0.0011	2.9e-05	38.629	0.000	0.001	0.00
per_pupil_spending	1.41e-08	1.44e-07	0.098	0.922	-2.68e-07	2.96e-0
avg_class_size	-5.944e-05	9.94e-05	-0.598	0.550	-0.000	0.00
asian_percent	0.0107	0.006	1.929	0.054	-0.000	0.02
black_percent	0.0107	0.006	1.920	0.055	-0.000	0.02
hispanic_percent	0.0107	0.006	1.926	0.055	-0.000	0.02
multiple_percent	0.0109	0.006	1.960	0.051	-2.96e-05	0.02
white_percent	0.0106	0.006	1.917	0.056	-0.000	0.02
rigorous_instruction	0.0001	0.001	0.151	0.880	-0.002	0.00
poverty_percent	-0.0002	4.8e-05	-4.250	0.000	-0.000	-0.00
ESL_percent	6.32e-05	4.34e-05	1.457	0.146	-2.2e-05	0.00
school_size	5.164e-06	2.19e-06	2.359	0.019	8.62e-07	9.47e-0
school_climate	0.0002	0.000	0.605	0.546	-0.000	0.00
appplication_rate	0.2074	0.014	14.958	0.000	0.180	0.23
poverty_percent_median	0.0029	0.001	2.497	0.013	0.001	0.00
over_spending_median	-0.0023	0.001	-2.069	0.039	-0.005	-0.00

Omnibus:	247.430	**Durbin-Watson:**	2.156
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	11919.115
Skew:	1.623	**Prob(JB):**	0.00
Kurtosis:	28.031	**Cond. No.**	3.22e+07

Finally, in terms of **acceptances_per_student**, we can see that there are 7 significant columns, 4 of which don't carry any information on application numbers, unlike applications and acceptances, which we can ignore. However, if our alpha levels were a bit higher for the cutoff (ie. 0.1 instead of 0.05), we would have 12 significant columns! These include the various columns regarding race, signifying a statistical difference between the acceptance chances for students of different races.

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 school characteristics that better determine the acceptance of students into HSPHS is **poverty_percent,** and although it's statistical significance is over the alpha cutoff of 0.05, it is only above by around a centesimal point. **Over_spending_median** is also a significant factor, and, at first, I believed that these two factors were correlated, but they are not. We can choose these two factors as the two most important ones given that the other factors deemed significant are columns directly related to acceptances (applications, school_size, application_rate, and acceptances_per_student). More students in a school means more applicants and higher chances of entering HSPHS.

Interestingly, we can see that P-Values for columns regarding race percentages are very low, near a value of 0.1, meaning that these values may also be factors in determining a school's acceptances into HSPHS.

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.

For sending more students into HSPHS, the New York City Department of Education has to deal with the poverty disparities between high schools and the providing of funds to schools to better provide resources to each and every student. There are clear disparities between schools below and over **poverty and spending_per_pupil** medians that need to be addressed. Selective Bias is a major problem in colleges and universities, and although we wouldn't want this problem to spill into high schools, perhaps dealing with social structures in a better fashion will yield better results for the students of the future.

In terms of providing a better education and the ability to achieve higher achievement scores, by far the most important factors for this dependent variable are r**igorous_education**, **poverty_percentages**, **ESL_percent**, and **school_climate**. School rigour is easy to deal with, although it has to be normalized across all schools in the City to provide the same level of information to students. This goes along with the school climate, where some schools are perceived differently by their respective students. ESL percentage is also important for what was previously mentioned, because perhaps children of immigrants have a more negative perception of their school and are less inclined to follow coursework according to the city's department of education guidelines, yielding lower achievement scores. Poverty percentages is also important

given that richer high school students are afforded more resources, opportunities, and perhaps even better staff than those in poorer schools.

Link to code is here:

https://github.com/rodriq9890/Data-Analysis-Project