Education Inequality

DATA 3320 project

Lauren Louie

About

- Address the topic of education inequality across high schools in the United States by focusing on student academic performance through ACT scores.
- The purpose of this project is to answer the question of *“Can school performance be predicted by socioeconomic factors?*
- Additionally, I will be looking into how living in a multilingual household can affect one’s ACT score.

Data

EdGap

- Primary dataset
- Taken from 2016
- Information about average ACT or SAT scores

school_info

- From National Center for Education Statistics
- Basic identifying information about schools

census_df

- S1601 data from the US Census Bureau
- Helpful for the multilingual part of analysis
- Key elements we are interested in are: person's age 5 and older, persons who speak only English, persons who speak a language other than English, and persons who speak English "Less than very well."

Linear Regression

- A standard linear regression table is shown below where the results predict the 'average_act' score based on features such as 'median_income', 'rate_unemployment', 'percent_college', 'percent_lunch', and 'percent_married'.

OLS Regression Results

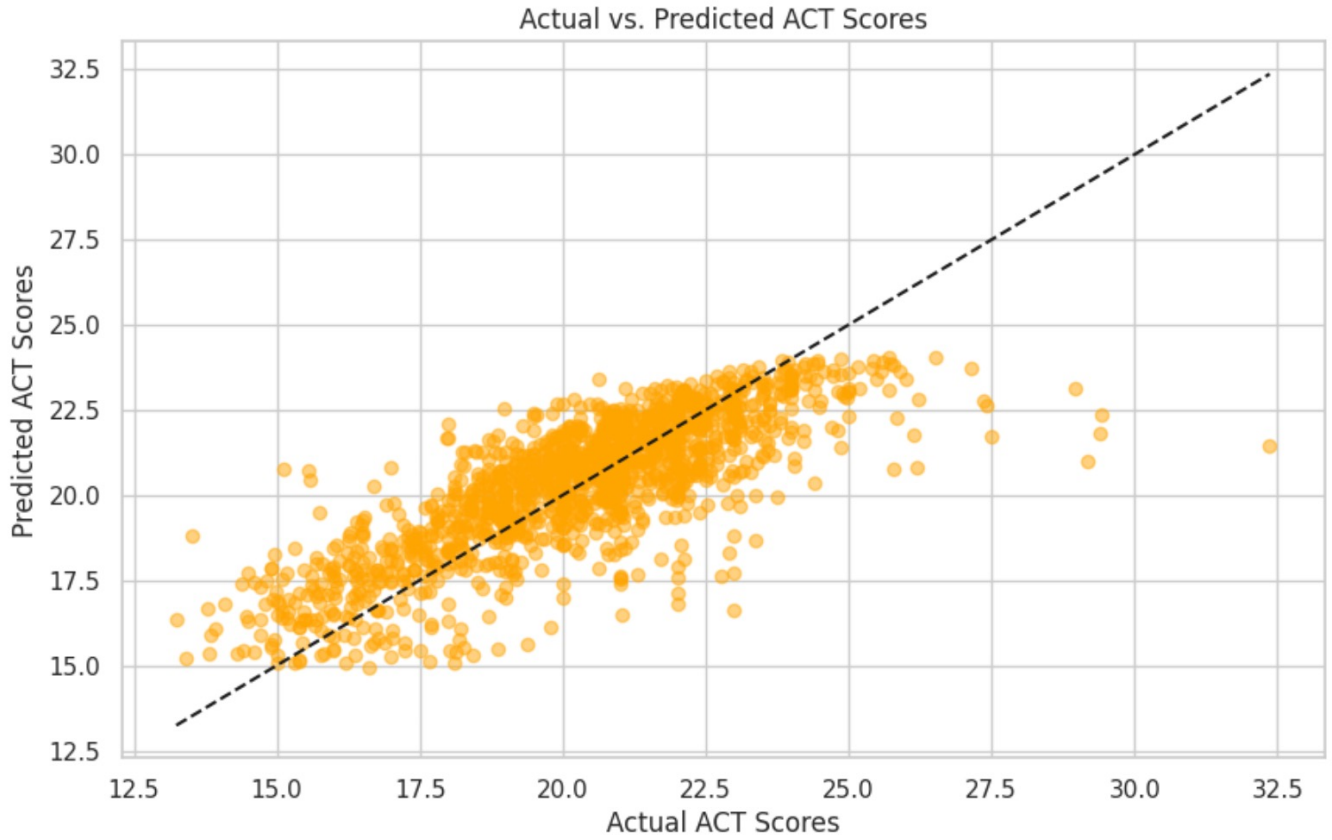
Dep. Variable:	average_act	R-squared:	0.641
Model:	OLS	Adj. R-squared:	0.641
Method:	Least Squares	F-statistic:	2534.
Date:	Tue, 07 May 2024	Prob (F-statistic):	0.00
Time:	07:33:27	Log-Likelihood:	-12973.
No. Observations:	7100	AIC:	2.596e+04
Df Residuals:	7094	BIC:	2.600e+04
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	20.2865	0.018	1135.920	0.000	20.252	20.322
median_income	-0.0166	0.029	-0.567	0.571	-0.074	0.041
rate_unemployment	-0.1235	0.023	-5.432	0.000	-0.168	-0.079
percent_college	0.2787	0.026	10.742	0.000	0.228	0.330
percent_lunch	-1.8193	0.023	-80.509	0.000	-1.864	-1.775
percent_married	-0.0220	0.025	-0.864	0.388	-0.072	0.028

Omnibus:	1002.935	Durbin-Watson:	2.017
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3073.508
Skew:	0.737	Prob(JB):	0.00
Kurtosis:	5.867	Cond. No.	3.57

Actual vs. Predicted ACT Scores

- The graph below shows uses both train and test model to calculate the predicted ACT scores based on socioeconomic factors and compared the to the actual average ACT scores.

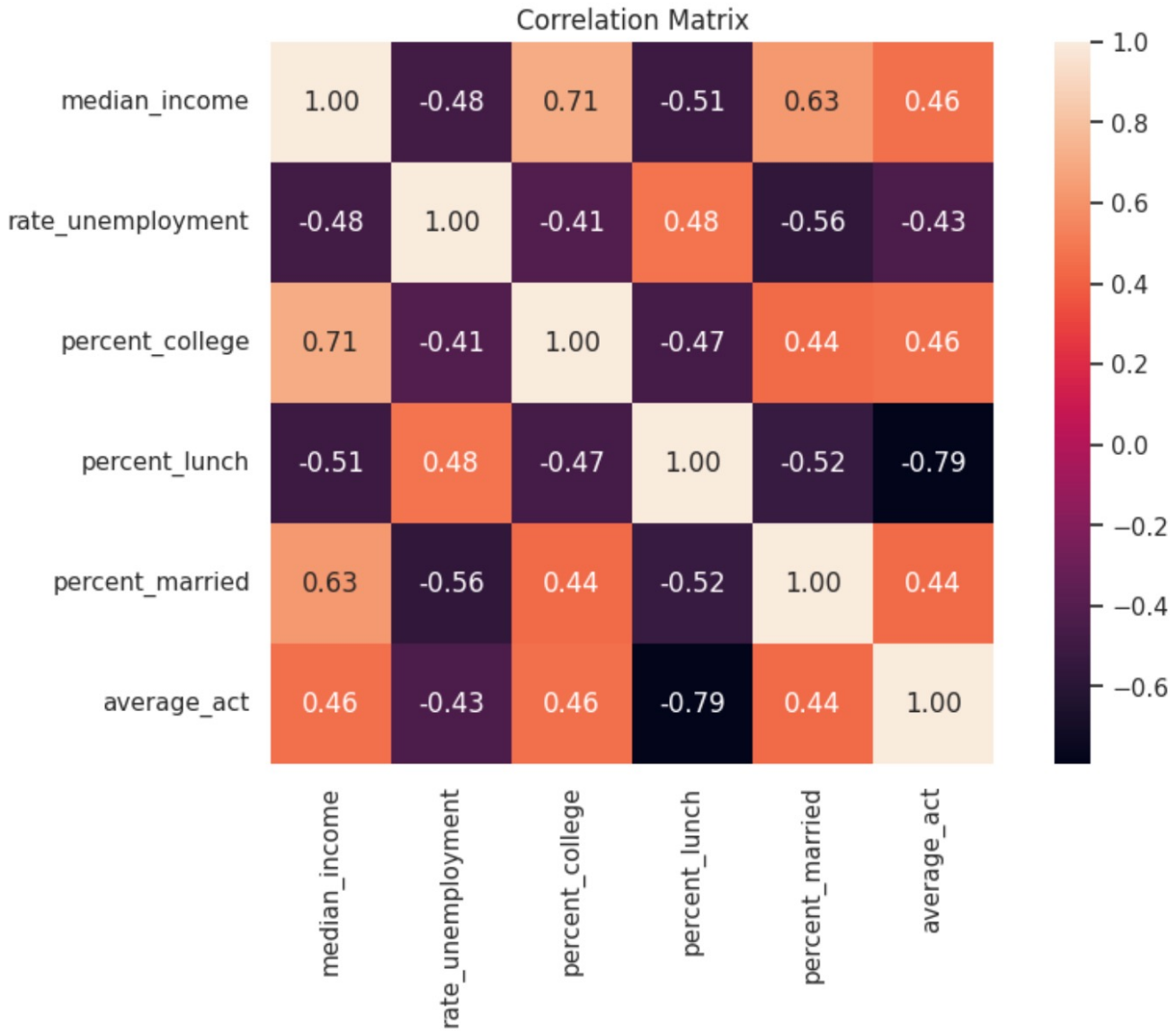


Questions to answer

1. *Which feature(s) has the highest correlation with average ACT score?*
2. *How do various socioeconomic factors influence ACT score?*
3. *What is the relationship between ACT score and multilingualism?*

Which feature(s) has the highest correlation with average ACT score?

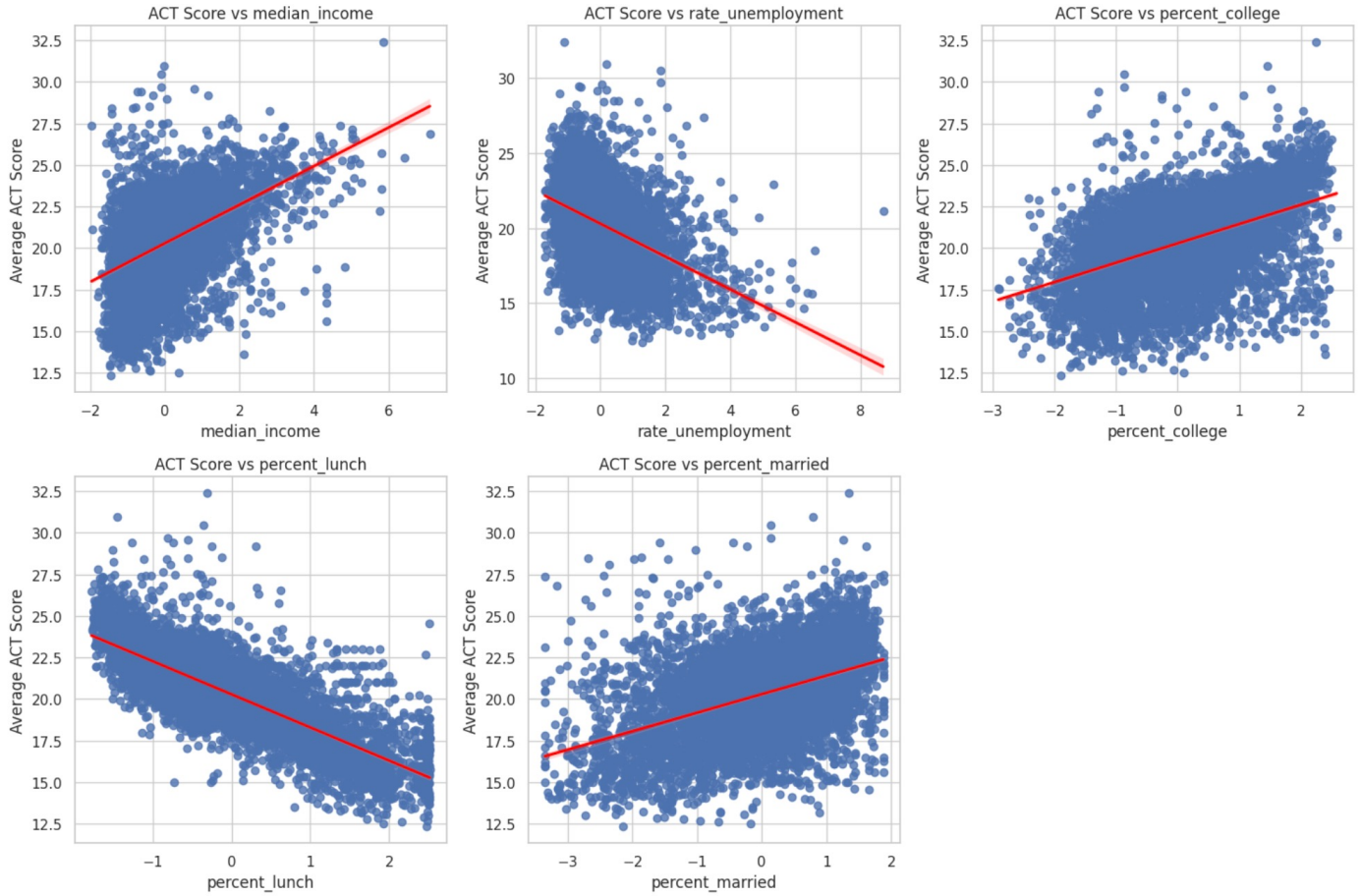
- The correlation matrix includes the correlations between features and target variable "average_act", providing insight into how each feature is related to the target variable.
- Besides average_act and average_act, the highest correlation is tied between **median_income** and **percent_college.**
- This can indicate a strong positive relationship between those variables and average_act compared to the other variables.



How do various socioeconomic factors influence ACT score?

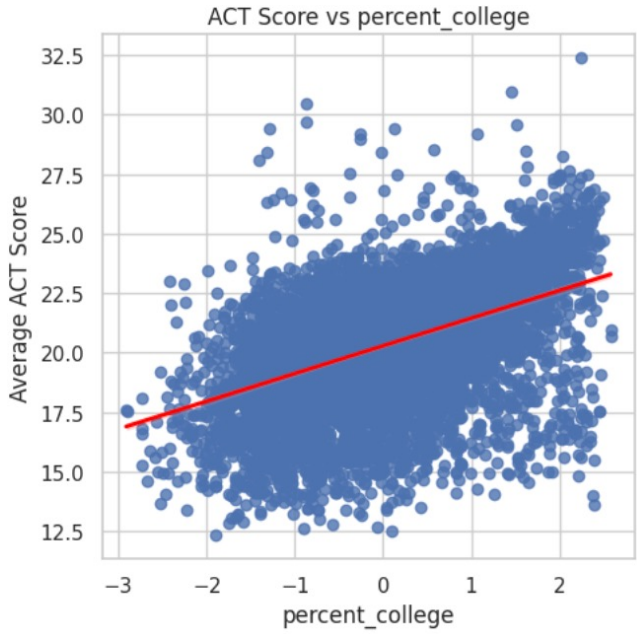
The five series of scatterplots for each of the variables that were initally kept in the beginning of the notebook.

Each include a line of best fit which helps indicate the correlation and linear equation for line of best fit.



How do various socioeconomic factors influence ACT score?

- Focusing on the two variables that tied for having the highest correlation, both show a positive relationship, yet have major outliers amongst the rest of the data.
- Both variables make logical sense as when a family's **median_income** is higher, they tend to have more disposable income that can be used for ACT tutors. Likewise, when your family has a greater **percent_college**, they tend to have higher paying jobs and can afford to spend money on expensive testing books or tutoring as well.



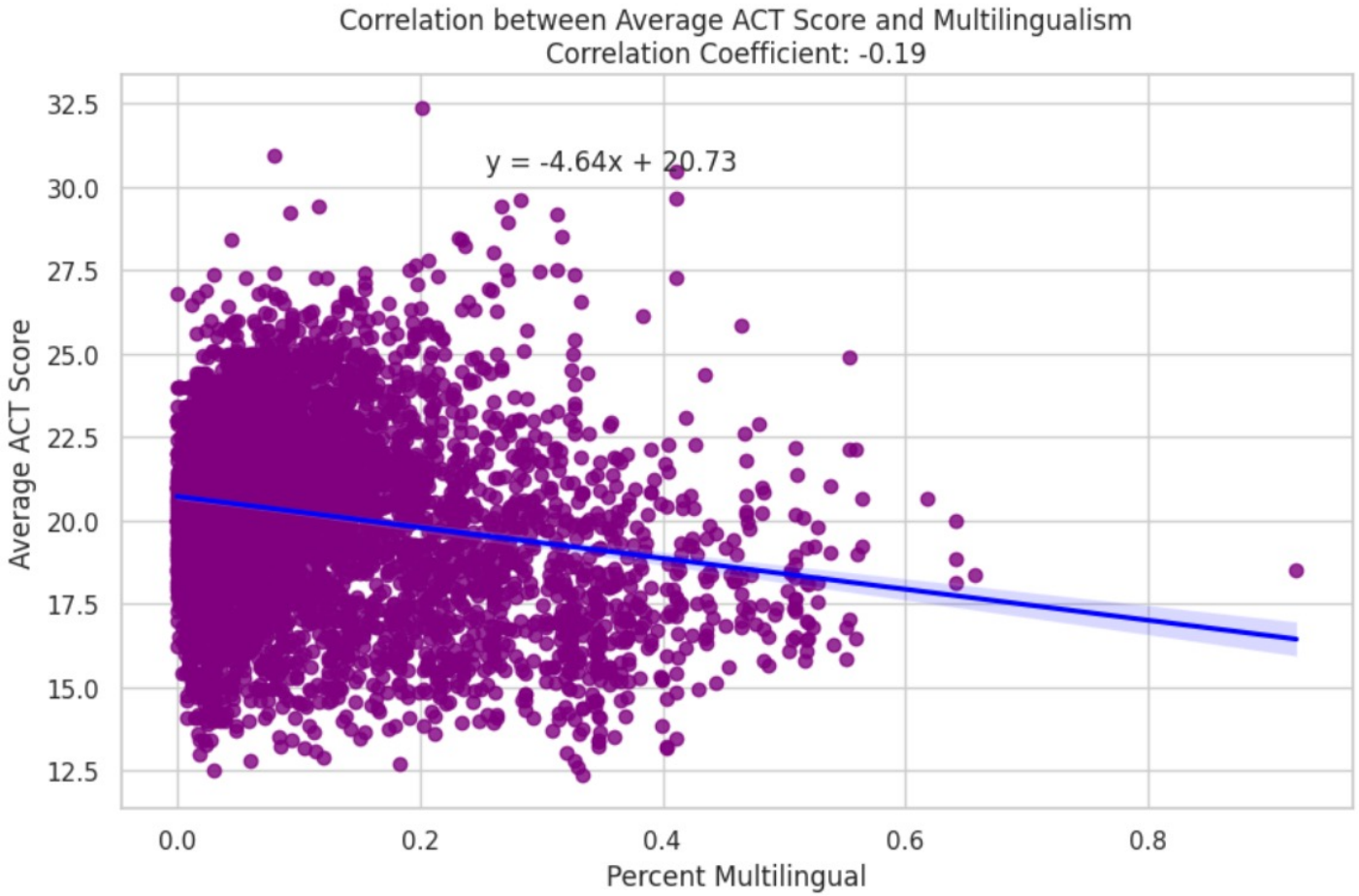
What is the correlation between average ACT score and multilingualism?

The code used to create a scatterplot and find the correlation value between the variables of average_act and percent_multilingual.

```
# Compute the correlation coefficient
correlation_coefficient = clean_df['average_act'].corr(clean_df['percent_multilingual'])

# Plot the scatter plot with regression line
plt.figure(figsize=(10, 6))
sns.regplot(x='percent_multilingual',
            y='average_act',
            color='purple',
            data=clean_df,
            line_kws={'color': 'blue'})  # Customize line color
plt.title(f'Correlation between Average ACT Score and Multilingualism\nCorrelation Coefficient: {correlation_coefficient:.2f}')
plt.xlabel('Percent Multilingual')
plt.ylabel('Average ACT Score')
plt.grid(True)

# Get regression line parameters
slope, intercept = np.polyfit(clean_df['percent_multilingual'], clean_df['average_act'], 1)

# Add equation of the regression line
equation_text = f'y = {slope:.2f}x + {intercept:.2f}'
plt.text(0.5, 0.9,
         equation_text,
         fontsize=12,
         ha='right',
         va='top',
         transform=plt.gca().transAxes)

plt.show()
```

What is the correlation between average ACT score and multilingualism?

- A -0.19 correlation coefficient indicates a weak negative correlation between the variables.
- It suggests that there is little to no linear relationship.



Conclusion

- Based on the visualizations and analysis completed, I can conclude that there is not a major difference between the percent of multilingual identifying population and average ACT scores based on the data given.
- However, there is an overall correlation between socioeconomic factors and ACT scores that we can see in our correlation matrix and other visualizations too.
- Moving forward, another interesting point that may be interesting is to use the number of attempts at the ACT and compare that to various socioeconomic factors.
- To answer the question of the project "*Can school performance be predicted by socioeconomic factors?*", the answer is **yes** it can.