Data Overview and Preprocessing

The dataset contains information on SAT scores aeducational and demographic factors across 51 observations.

The variables include:

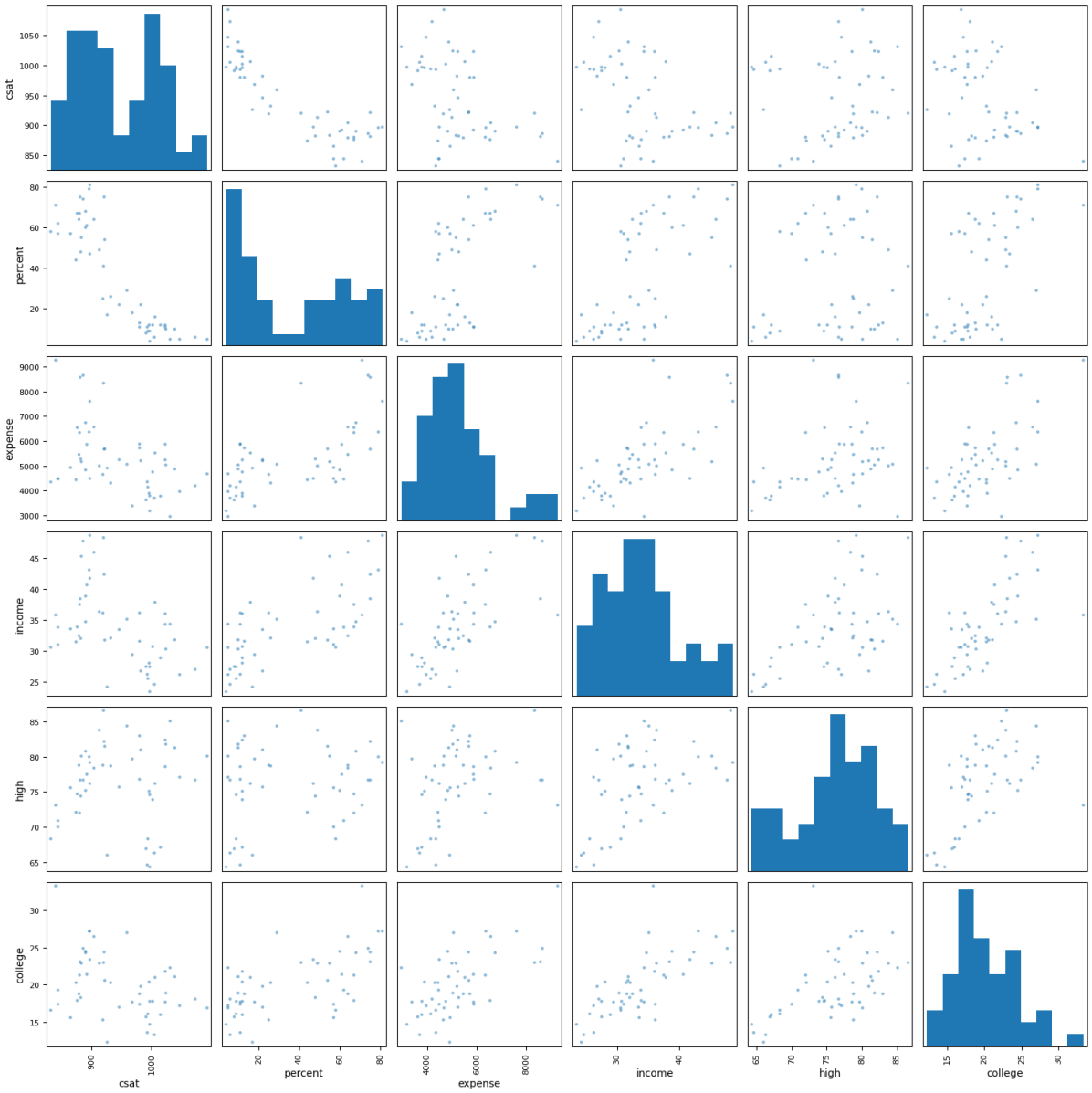
- region : Geographical region (categorical)
- csat : Mean composite SAT score (dependent variable)
- percent : Percentage of high school graduates taking the SAT
- expense : Per-pupil expenditures in primary and secondary education
- income : Median household income (in thousands of dollars)
- high : Percentage of adults with a high school diploma
- college : Percentage of adults with a college degree

The 'region' variable was initially categorical with one missing value. To incorporate it into our regression model, I encoded it numerically: West=1, N. East=2, South=3, Midwest=4, and assigned 9 to the missing value. This encoding preserves the categorical nature while allowing its use in regression analysis.

The distribution of regions (South: 16, West: 13, Midwest: 12, N. East: 9, Missing: 1) shows a slight imbalance, with the South being the most represented region. This imbalance should be kept in mind when interpreting results, as it may affect the generalizability of our findings to the broader United States

Correlation Matrix

Variable	Region	CSAT	Percent	Expense	Income	High	College
Region	1	0.1248	-0.1227	0.1499	-0.1999	-0.3432	0.0092
CSAT	0.1248	1	-0.8758	-0.4663	-0.4713	0.0858	-0.3729
Percent	-0.1227	-0.8758	1	0.6509	0.6733	0.1413	0.6091
Expense	0.1499	-0.4663	0.6509	1	0.6784	0.3133	0.6399
Income	-0.1999	-0.4713	0.6733	0.6784	1	0.5099	0.7233
High	-0.3432	0.0858	0.1413	0.3133	0.5099	1	0.5319
College	0.0092	-0.3729	0.6091	0.6399	0.7233	0.5319	1



- CSAT scores show strong negative correlations with percent (-0.876) and expense (-0.466). This suggests that as the percentage of students taking the SAT increases, or as per-pupil expenditures increase, SAT scores tend to decrease. The negative correlation with percent might be explained by a broader, more diverse pool of test-takers leading to lower average scores. The negative correlation with expense is more surprising and warrants further investigation.

- There's a moderate positive correlation between high school completion rates (high) and college attendance rates (college) (0.532). This is expected, as areas with higher high school completion rates are likely to have more students continuing to higher education.

- Income shows strong positive correlations with expense (0.678) and college (0.723). This indicates that higher-income areas tend to spend more on education and have higher college attendance rates. This aligns with common socioeconomic patterns where wealthier areas invest more in education and have higher rates of college attendance.

- The region variable shows weak correlations with most other variables, with the strongest being a negative correlation with high school completion rates (-0.343). This suggests that regional differences may not be as pronounced as other factors in our dataset, but there might be some regional variation in high school completion rates.

These correlations provide a preliminary understanding of the relationships in our data, but they don't account for the simultaneous effects of multiple variables. For this, I turn to our multiple regression model.

Multiple Linear Regression Results

Dep. Variable:	csat	**R-squared:**	0.826
Model:	OLS	**Adj. R-squared:**	0.803
Method:	Least Squares	**F-statistic:**	34.93
Date:	Sat, 21 Sep 2024	**Prob (F-statistic):**	3.54e-15
Time:	17:52:46	**Log-Likelihood:**	-241.59
No. Observations:	51	**AIC:**	497.2
Df Residuals:	44	**BIC:**	510.7
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	823.2891	70.910	11.610	0.000	680.379	966.199
region	2.7464	3.732	0.736	0.466	-4.775	10.268
expense	0.0018	0.005	0.371	0.713	-0.008	0.012
percent	-2.5478	0.273	-9.350	0.000	-3.097	-1.999
income	0.2961	1.200	0.247	0.806	-2.123	2.715
high	2.0052	1.120	1.791	0.080	-0.251	4.262
college	1.6001	1.768	0.905	0.370	-1.964	5.164

Omnibus:	0.851	**Durbin-Watson:**	2.239
Prob(Omnibus):	0.653	**Jarque-Bera (JB):**	0.939
Skew:	0.255	**Prob(JB):**	0.625
Kurtosis:	2.575	**Cond. No.**	9.23e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.23e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Our multiple linear regression model attempts to explain the variation in SAT scores using all other variables as predictors.

The R-squared value of 0.826 indicates that our model explains 82.6% of the variance in CSAT scores. The adjusted R-squared of 0.803 suggests that this explanatory power holds even when penalizing for the number of predictors. This indicates a good overall fit, suggesting our chosen variables capture a substantial portion of what influences SAT scores.

The F-statistic of 34.93 with a p-value of 3.54e-15 indicates that our model as a whole is statistically significant. I can reject the null hypothesis that all our coefficient values are zero, confirming that our predictors, taken together, have a meaningful relationship with SAT scores.

Coefficients:

- For each unit increase in the region code, the SAT score is expected to increase by 2.7464 points, holding other variables constant. However, this effect is not statistically significant (p-value = 0.466). This suggests that after accounting for other factors, regional differences may not play a significant role in determining SAT scores.

- A one-unit (dollar) increase in per-pupil expenditures is associated with a 0.0018 point increase in SAT scores, all else being equal. This effect is not statistically significant (p-value = 0.713). The small coefficient and lack of significance suggest that, contrary to what I might expect, increasing educational expenditures may not directly translate to higher SAT scores when other factors are considered.

- This variable has the most substantial and statistically significant effect. For each percentage point increase in students taking the SAT, the score is expected to decrease by 2.5478 points (p-value < 0.001). This strong negative relationship might be explained by the fact that as more students take the test, a wider range of academic abilities is represented, potentially lowering the average score.

- A one-unit ($1000) increase in median household income is associated with a 0.2961 point increase in SAT scores, but this effect is not statistically significant (p-value = 0.806). The positive coefficient aligns with expectations that higher income might be associated with better educational outcomes, but the lack of significance suggests this relationship may be captured by other variables in the model.

- For each percentage point increase in adults with a high school diploma, the SAT score is expected to increase by 2.0052 points. This effect is marginally significant (p-value = 0.080). This suggests that areas with higher high school completion rates tend to have slightly higher SAT scores, possibly reflecting a generally stronger educational environment.

- A one percentage point increase in adults with a college degree is associated with a 1.6001 point increase in SAT scores, but this effect is not statistically significant (p-value = 0.370). The positive coefficient aligns with expectations, but the lack of significance might be due to collinearity with other variables like income or high school completion rates.

The condition number of 9.23e+04 is quite large, suggesting potential multicollinearity or other numerical problems. This indicates that while our model provides a good fit, there may be instability in our coefficient estimates due to correlations among our predictors.

Homoskedasticity Test

Breusch-Pagan test p-value : 0.473689599112849

The Breusch-Pagan test yielded a p-value of 0.4737, which is greater than our typical significance level of 0.05. This means I fail to reject the null hypothesis of homoskedasticity. In practical terms, this suggests that the variance of our residuals is relatively constant across different levels of our predicted values. This is a desirable property for our linear regression, as it supports the reliability of our standard errors and, consequently, our t-tests and F-tests.

Multicollinearity Check

Feature	VIF
const	290.2651
region	1.6292
expense	2.7177
percent	2.8832

income	3.3641
high	2.2158
college	3.0712

Variance Inflation Factors (VIF) help us assess multicollinearity in our model:

- All VIF values for our predictors are below 4, which is generally considered acceptable. This suggests that multicollinearity is not a severe issue in our model.
- The highest VIF is for income (3.364), followed by college (3.071) and percent (2.883). While these are not alarmingly high, they do indicate some correlation among these predictors. This might explain why income and college, despite showing strong correlations with SAT scores in our correlation matrix, are not significant in our regression model.
- The low VIF for region (1.629) suggests that it provides relatively independent information compared to our other predictors.

These VIF values, while not indicating severe multicollinearity, do suggest some interdependence among our predictors. This aligns with our earlier observation of correlations among variables like income, college, and expense.

Normality Test

Shapiro-Wilk test p-value : 0.43283144171363697

The Shapiro-Wilk test for normality of residuals yielded a p-value of 0.433, which is greater than 0.05. This means I fail to reject the null hypothesis that our residuals are normally distributed. This supports the assumption of normality in our regression model, which is important for the validity of our t-tests and F-tests. It suggests that our model's errors are reasonably well-behaved, adding credibility to our inference from the model.

F-test

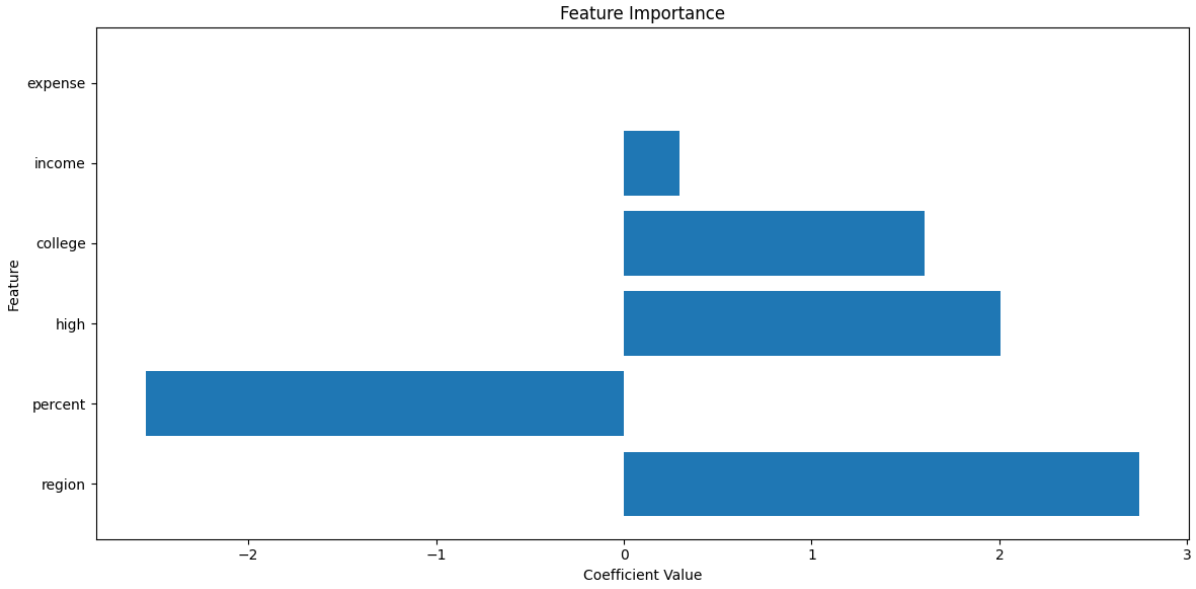
F-test Results :

<F test: F=3.5365426750623516, p=0.03764721044530699, df_denom=44, df_num=2>

I conducted a F-test for the joint hypothesis that the coefficients of 'high' and 'college' are both zero. This test yielded an F-statistic of 3.5365 with a p-value of 0.0376, which is less than 0.05. This allows us to reject the null hypothesis at the 5% significance level.

Even though 'high' and 'college' weren't individually significant at the 0.05 level in our regression output, this F-test suggests that together, they have a statistically significant effect on SAT scores. This highlights the importance of considering joint effects in our model, rather than focusing solely on individual coefficient tests.

Feature Importance



Our feature importance plot visualizes the magnitude and direction of each predictor's effect on SAT scores:

- Percent has the largest absolute coefficient, indicating it has the strongest influence on SAT scores, with a negative relationship. This aligns with our regression results and

suggests that the proportion of students taking the SAT is the most crucial factor in determining average scores.

- High school completion rate (high) and college attendance rate (college) have the next largest positive coefficients. This suggests they have substantial positive associations with SAT scores, even though college wasn't statistically significant in our regression output.
- Region and income show smaller positive effects. The small effect of region aligns with its non-significance in our regression model. The small effect of income, despite its strong correlations with other variables, might be due to its effects being partially captured by other variables in the model.
- Expense has a very small positive coefficient, barely visible on the plot. This aligns with its non-significant p-value in the regression output and suggests that, in our model, changes in educational expenses have minimal direct impact on SAT scores when controlling for other factors.

The coefficient for expense (0.001833) is very small, which explains why it's barely visible in our feature importance plot. This suggests that, in our model, changes in educational expenses have minimal direct impact on CSAT scores when controlling for other factors. This is a surprising finding that goes against common assumptions about the impact of educational spending on outcomes.