Python Analytics for Student Performance Management

Inspiration & Problem

Inspiration:

Educators often lack intuitive tools to quickly assess which students are thriving and which may need support. By analyzing a dataset, we aim to understand the effect of different factors (like test prep or parental education) on student outcomes, and create visual dashboards to help educators identify and support at-risk students.

Problem:

We want to identify the effects of student involvement by creating a student management system incorporated with various attributes such as gender and race/ethnicity, parental level of education, test preparation, and performance score of math, reading, and writing.

Background of Dataset

- We have used the Student Performance Dataset
- The dataset from GitHub, created by @rashida048.
- The dataset contains data from 1000 students and includes demographic details (gender, race/ethnicity, parental education), test preparation status, lunch type, and performance scores in math, reading, and writing.

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group B	bachelor's degree	standard	none	72	72	74
1	female	group C	some college	standard	completed	69	90	88
2	female	group B	master's degree	standard	none	90	95	93
3	male	group A	associate's degree	free/reduced	none	47	57	44
4	male	group C	some college	standard	none	76	78	75
5	female	group B	associate's degree	standard	none	71	83	78
6	female	group B	some college	standard	completed	88	95	92
7	male	group B	some college	free/reduced	none	40	43	39
8	male	group D	high school	free/reduced	completed	64	64	67
9	female	group B	high school	free/reduced	none	38	60	50
10	male	group C	associate's degree	standard	none	58	54	52
11	male	group D	associate's degree	standard	none	40	52	43
12	female	group B	high school	standard	none	65	81	73
13	male	group A	some college	standard	completed	78	72	70
14	female	group A	master's degree	standard	none	50	53	58

Questions

- Does gender influence the level of student involvement and academic performance across core subjects?
- Is there a relationship between parental education levels and student academic achievement?
- Do students from different racial or ethnic backgrounds show significant differences in academic performance?

- Among the top 10 highest-performing students (based on average scores), what common attributes can be identified (e.g., gender, race/ethnicity, parental education)?
- How does completion of a test preparation course impact student scores?
- Which gender performs better in each subject: math, reading, and writing?

Plan

- Class: Our class would be the Student, and each object of this class represents a single student from the data set
- Attributes :
  - Gender -female or male
  - Race_ethnicity- group A, group B
  - Parental_education - bachelor's degree
  - Test_prep - Completed or none
  - Math_score, reading_score, and writing_score

	0
gender	0
race/ethnicity	0
parental_level_of_education	0
lunch	0
test_preparation_course	0
math_score	0
reading_score	0
writing_score	0

Objective

We expect the Student Management System to provide insightful information and data-driven conclusions about student performance trends, and the impact of various factors when it comes to their academic performance. The goal is to help educators make informed decisions by finding key patterns and finding clear visualizations within the management system.

Dataset Overview

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 9 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   gender                       1000 non-null   object
 1   race/ethnicity               1000 non-null   object
 2   parental_level_of_education  1000 non-null   object
 3   lunch                        1000 non-null   object
 4   test_preparation_course      1000 non-null   object
 5   math_score                   1000 non-null   int64
 6   reading_score                1000 non-null   int64
 7   writing_score                1000 non-null   int64
 8   average_score                1000 non-null   float64
dtypes: float64(1), int64(3), object(5)
memory usage: 70.4+ KB
```

.info

	math_score	reading_score	writing_score	average_score
count	1000.00000	1000.000000	1000.000000	1000.000000
mean	66.08900	69.169000	68.054000	67.770667
std	15.16308	14.600192	15.195657	14.257326
min	0.00000	17.000000	10.000000	9.000000
25%	57.00000	59.000000	57.750000	58.333333
50%	66.00000	70.000000	69.000000	68.333333
75%	77.00000	79.000000	79.000000	77.666667
max	100.00000	100.000000	100.000000	100.000000

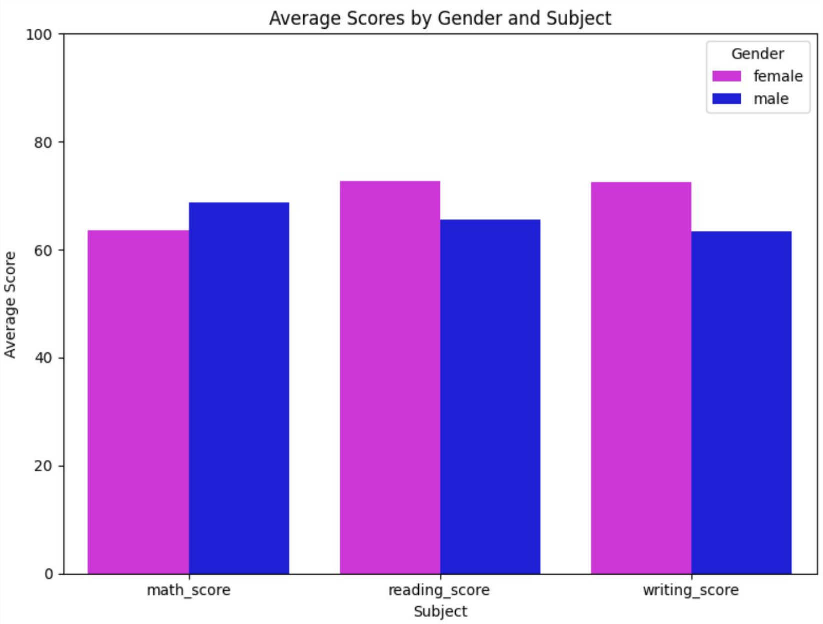
.describe

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
995	female	group E	master's degree	standard	completed	88	99	95
996	male	group C	high school	free/reduced	none	62	55	55
997	female	group C	high school	free/reduced	completed	59	71	65
998	female	group D	some college	standard	completed	68	78	77
999	female	group D	some college	free/reduced	none	77	86	86

Avg Scores by Gender & Subject

Based on the bar plot comparing average scores by gender and subject:

- **Math**:
  - **Males** perform better than females.
  - Male average ≈ 69, Female average ≈ 64.
- **Reading**:
  - **Females** perform better than males.
  - Female average ≈ 73, Male average ≈ 66.
- **Writing**:
  - **Females** also outperform males.
  - Female average ≈ 72, Male average ≈ 64.



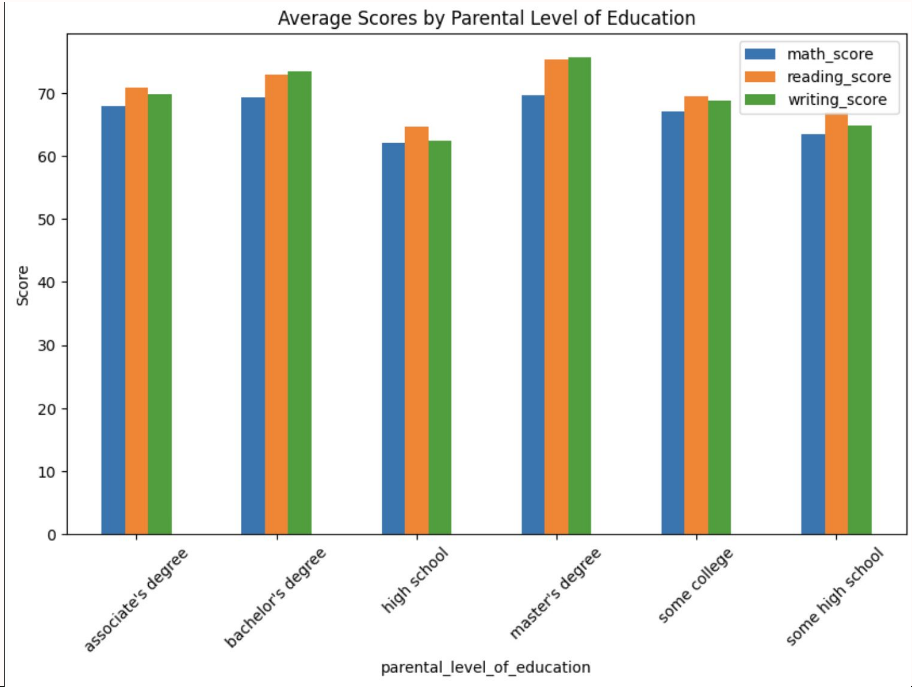

```python
avg_scores = data.groupby('gender')[['math_score', 'reading_score', 'writing_score']].mean().reset_index()
avg_scores_long = avg_scores.melt(id_vars='gender',
                                  var_name='Subject',
                                  value_name='Average Score')


plt.figure(figsize=(8, 6))
sns.barplot(data=avg_scores_long, x='Subject', y='Average Score', hue='gender',
            palette={'female': 'magenta', 'male': 'blue'})

plt.title('Average Scores by Gender and Subject')
plt.ylabel('Average Score')
plt.xlabel('Subject')
plt.ylim(0, 100)
plt.legend(title='Gender')
plt.tight_layout()
plt.show()
```

Parental Education vs Academic Performance

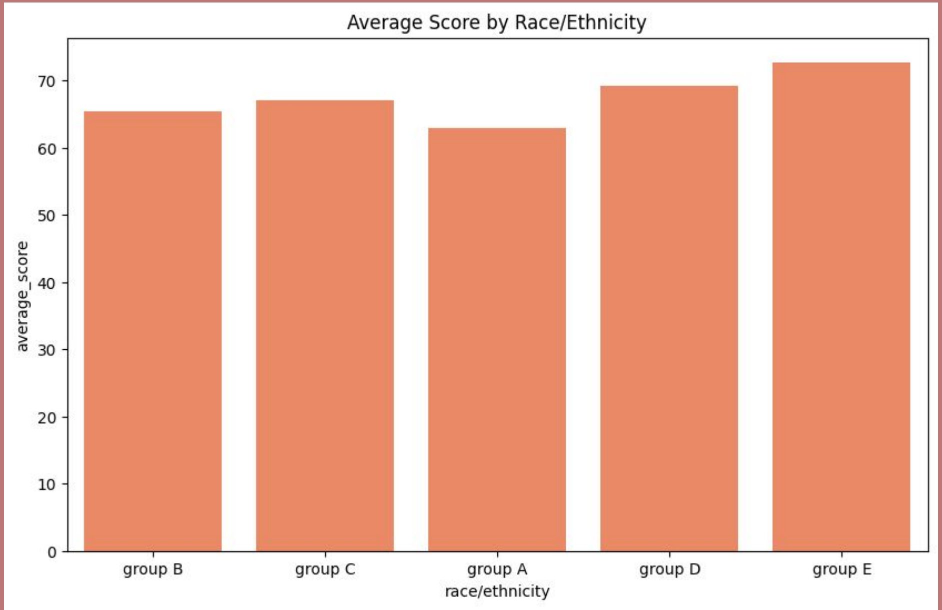
- Students with more highly educated parents (especially with master's or bachelor's degrees) score higher in all subjects.
- The highest scores are seen in students whose parents have a master's degree.
- The lowest scores are among those whose parents have only some high school or high school education.
- Reading and writing scores are consistently higher than math scores.
- Overall, higher parental education correlates with better academic performance.




```
# group and plot average scores by parental education
parental_avg = data.groupby('parental_level_of_education')[['math_score', 'reading_score', 'writing_score']].mean()
parental_avg.plot(kind='bar', figsize=(10,6))
plt.title("Average Scores by Parental Level of Education")
plt.ylabel("Score")
plt.xticks(rotation=45)
plt.show()
```

Academic Performance Across Race/Ethnicity

- Group E has the highest average academic score, followed by Group D.
- Group A shows the lowest average score among all groups.
- Overall, there is variation in performance across different race/ethnicity groups.
- This suggests that demographic factors may influence academic outcomes, possibly due to social, economic, or educational disparities.

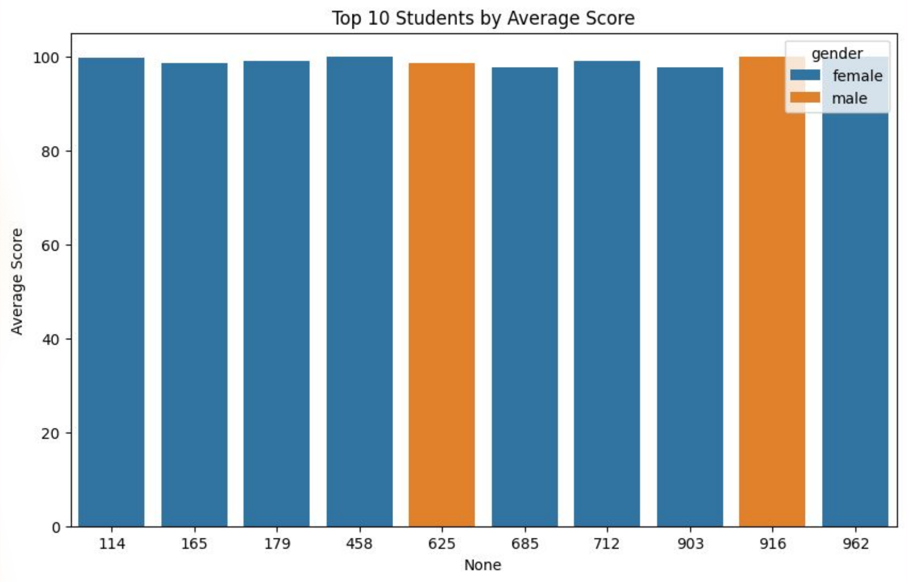



```
# Academic Performance Across Racial/Ethnic Groups
plt.figure(figsize=(10, 6))
sns.barplot(data=data, x='race/ethnicity', y='average_score', ci=None, color='coral')
plt.title('Average Score by Race/Ethnicity')
plt.show()
```

Top 10 Students (Gender Analysis)

Based on the bar plot, among the top 10 highest-performing students by average score:

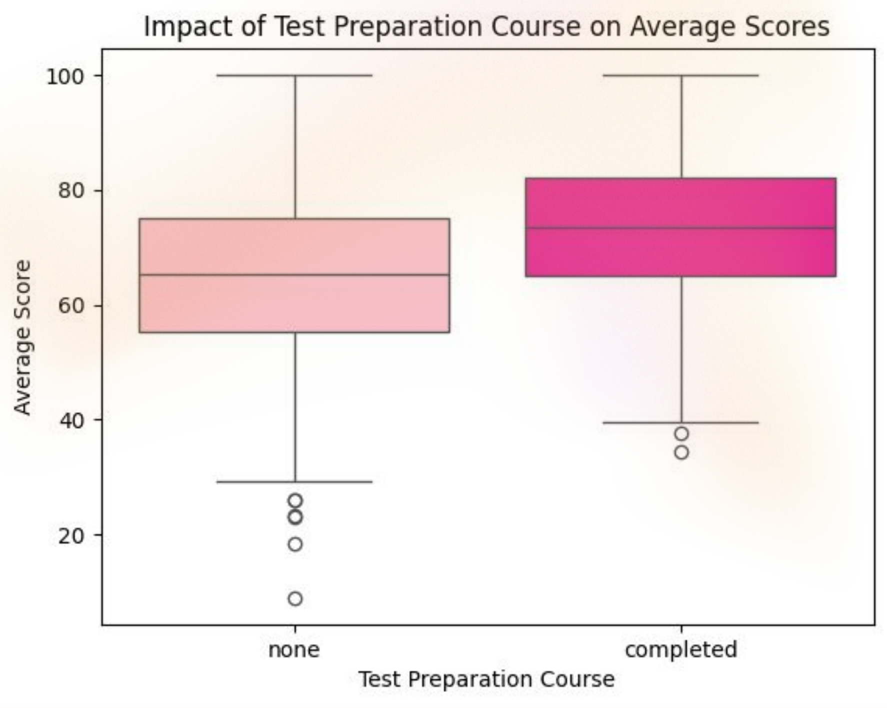
- **Most are female**, indicating a gender trend in top performance.
- **Only two are male**, suggesting females are more dominant among high scorers in this dataset.
- Although race/ethnicity and parental education are not shown directly in the chart, the strong female representation stands out as the most notable pattern.



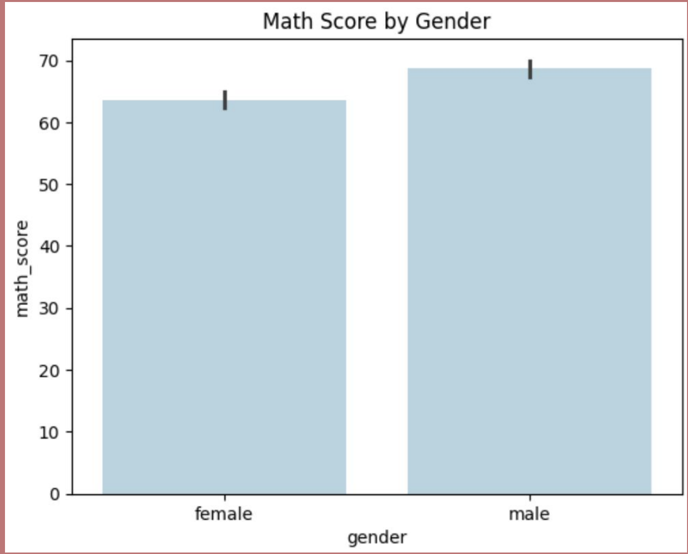

```
plt.figure(figsize=(10, 6))
sns.barplot(data=top_10, x=top_10.index, y='average_score', hue='gender')
plt.title('Top 10 Students by Average Score')
plt.ylabel('Average Score')
plt.show()
```

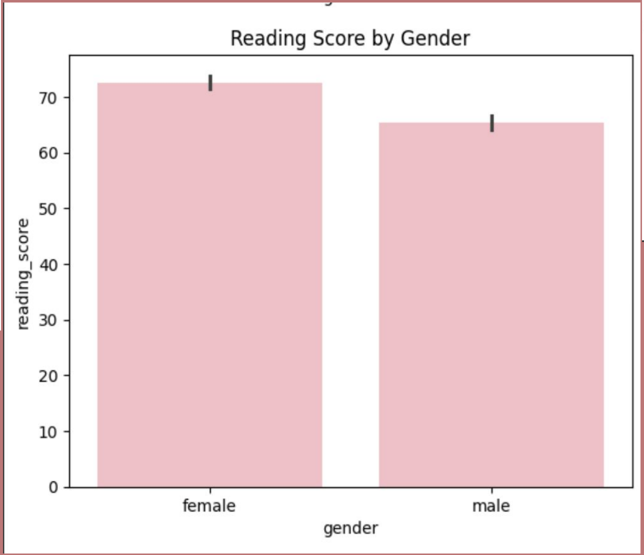
Impact of Test Preparation Course on Student Scores

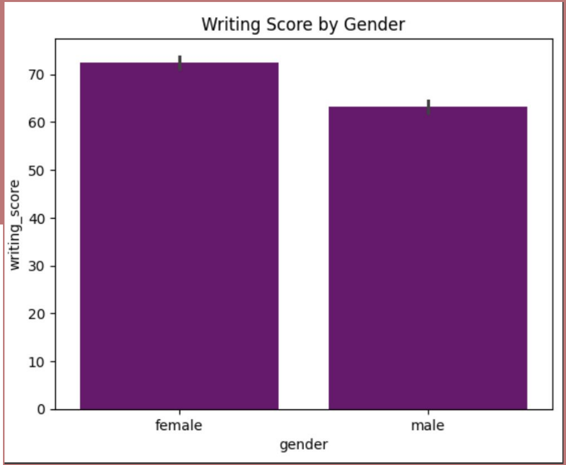
- Students who completed test prep scored higher on average.
- The median score is significantly higher for the completed group.
- Lower-end scores are much improved after test prep. Score variability is reduced—fewer low outliers.
- Top scores are similar for both groups, but the overall shift favors those who prepared.
- This suggests test preparation helps boost performance, especially for lower-performing students.



Gender Performance Comparison Across Academic Subjects: Math, Reading, and Writing








```python
sns.barplot(x='gender', y='math_score', data=data, color='lightblue')
plt.title("Math Score by Gender")
plt.show()

# Reading Score by Gender
sns.barplot(x='gender', y='reading_score', data=data, color='lightpink')
plt.title("Reading Score by Gender")
plt.show()

# Writing Score by Gender
sns.barplot(x='gender', y='writing_score', data=data, color='purple')
plt.title("Writing Score by Gender")
plt.show()
```

Statistical Significance

1. Gender vs. Math Score – Independent t-test

- The `ttest_ind()` function tests whether the means of two independent samples are significantly different.

```
from scipy.stats import ttest_ind

t_stat, p_value = ttest_ind(
    data[data['gender'] == 'male']['math_score'],
    data[data['gender'] == 'female']['math_score']
)
print(f"P-value: {p_value:.4f}")
```

```
P-value: 0.0000
```

Result:

- The **P-value is 0.0000**, which suggests:
    - There is a **statistically significant difference** between male and female math scores.
    - Since the P-value is extremely low (typically compared against α = 0.05), we **reject the null hypothesis** (which states that there is no difference between the means).

Statistical Significance

Gender vs. Reading and Writing Scores (t-test):

- A two-sample t-test was conducted to compare reading and writing scores between male and female students.
- **Reading Score:**
    - *t(998) = -9.98*, **p < 0.0001**
    - There is a **statistically significant difference** in reading scores between genders.

t-statistic: -9.98 → strong difference between groups.

p-value: 2.02e-22 → extremely small, **statistically significant difference** between male and female scores.

```
[37] # Reading
     ttest_ind(
         data[data['gender'] == 'male']['reading_score'],
         data[data['gender'] == 'female']['reading_score']
     )

     # Writing
     ttest_ind(
         data[data['gender'] == 'male']['writing_score'],
         data[data['gender'] == 'female']['writing_score']
     )

TtestResult(statistic=np.float64(-9.979557910004507), pvalue=np.float64(2.019877706867934e-22), df=np.float64(998.0))
```

Statistical Significance

Race/Ethnicity vs. Math Scores (ANOVA):

- **one-way ANOVA (Analysis of Variance)** test to examine whether there are **statistically significant differences** in **math scores** among different **race/ethnicity groups** in the dataset.

- **F-statistic**: 14.593885166332635 — a measure of the variance **between groups** relative to the variance **within groups**.
- **p-value**: 1.3732194030370688e-11 — an extremely small value, far less than common alpha levels (e.g., 0.05 or 0.01).

- Since the p-value is extremely low ($\approx 1.37 \times 10^{-11}$), you can **reject the null hypothesis**, which assumes that all race/ethnicity groups have the **same average math score**. This suggests that **at least one group's average math score is significantly different** from the others.

```
from scipy.stats import f_oneway

f_oneway(
    *[group['math_score'].values for name, group in data.groupby('race/ethnicity')]
)
```

```
F_onewayResult(statistic=np.float64(14.593885166332635), pvalue=np.float64(1.3732194030370688e-11))
```

Challenges/Limitations

- **Imbalanced Groups**:
  The dataset includes categorical variables like gender, race/ethnicity, and parental education.
  If one group (e.g., males or a specific race/ethnicity) dominates, it could bias conclusions.

- **Small Samples**:
  The dataset comes from a publicly available source, but categories with few members can make statistical analysis less reliable.

- **Data Cleaning Potential**:
  The code would have more potential if the dataset had a lot of missing or incorrect values to clean and handle.

Challenges/Limitations

Visualization difficulties, ambiguous trends, or anything unexpected

Uneven Distribution Across Groups (Race/Ethnicity vs. Parental Education):

- Some ethnicity groups (ex Group C) show a heavier concentration of parents with higher education while others (ex Group A) have more parents with lower education levels.
- This uneven spread might affect academic performance comparisons across groups.

Strong Score Correlations (Heatmap):

- There's a very strong positive correlation between reading and writing scores (0.95), suggesting they may be measuring similar abilities or influenced by similar factors.
- Math score correlations are also strong but slightly lower, indicating it might be somewhat independent compared to reading/writing.

Potential Confounding: Parental education and ethnicity could confound interpretations of academic performance, especially if not controlled for in analysis.

Tools and Techniques

- **Pandas**: Used for loading, cleaning, and transforming the dataset efficiently, including creating new columns and handling column names.

- **Seaborn**: Helped create high-level visualizations like the stacked bar chart and correlation heatmap with minimal code, making trends easier to spot.

- **Matplotlib**: Supported fine-tuning and customizing plots, such as adjusting figure size and labels, which improved clarity.

Based on the Data...

- **Encourage broader access to test prep courses**: Students who completed test preparation tend to perform better across subjects. Expanding access can help close performance gaps.

- **Provide targeted support to students with less-educated parents**: The visualizations show that parental education levels vary by race/ethnicity, and students with less-educated parents may be at a disadvantage. Support programs can help bridge this gap.

- **Address racial/ethnic achievement gaps through tailored interventions**: There are clear differences in parental education across racial/ethnic groups, which may contribute to unequal academic outcomes. Interventions should be culturally aware and specifically designed to support the needs of each group.