# Investigating the Factors Influencing Academic Performance in Students

Objective: To analyze and understand the various factors that contribute to the differences in academic performance among students, focusing on key aspects such as study habits, parental education levels, and other socio-economic and personal variables.

Background: The school has observed a significant variance in students' academic marks and seeks to identify the underlying reasons for this disparity. Understanding these factors is crucial for implementing targeted interventions and support mechanisms to enhance student performance across the board.
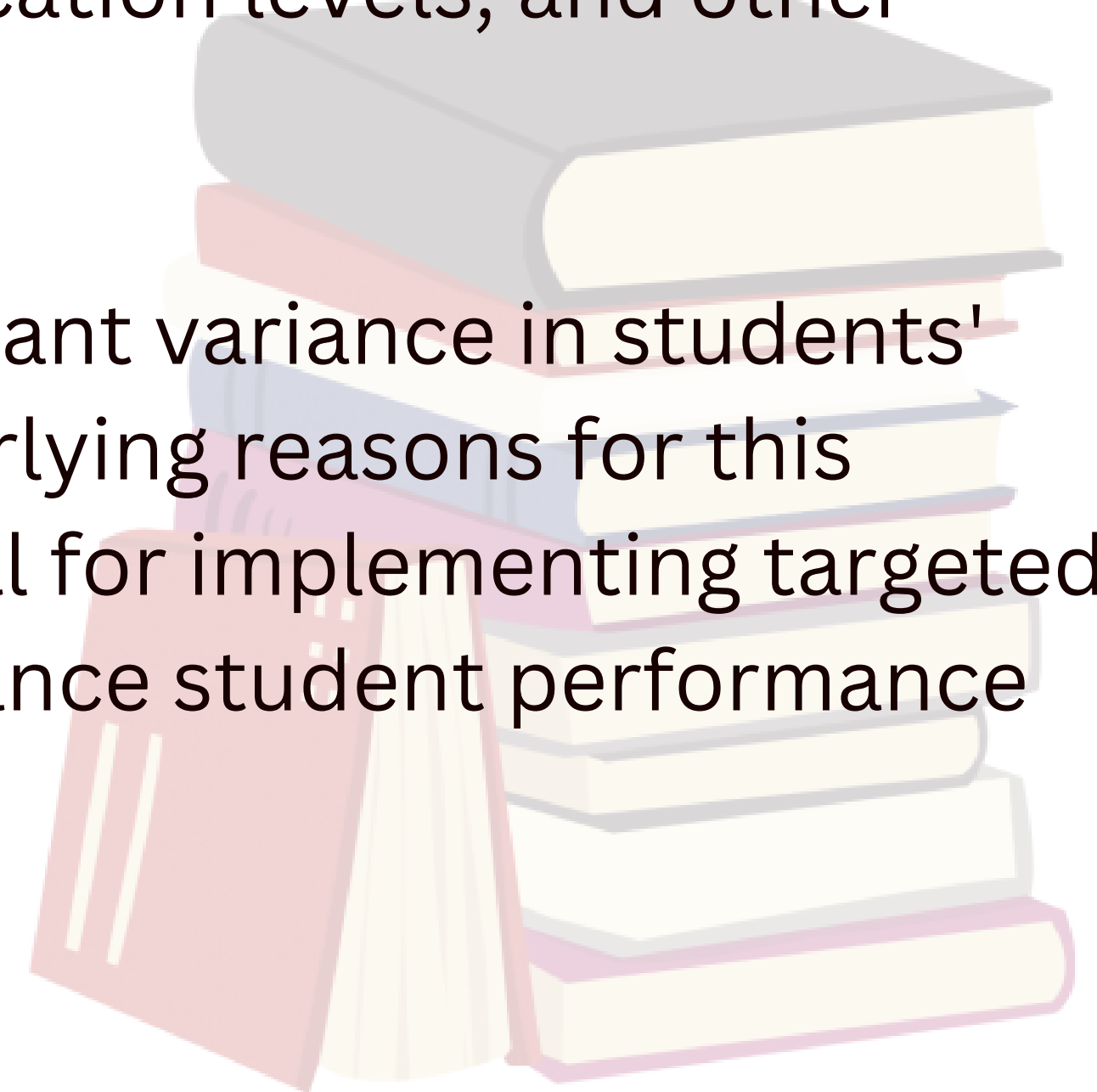
Key Points of Investigation:

1. **Study Habits:**
   - Number of hours dedicated to study weekly or daily.
   - Study environment and resources if available.
   - Participation in extracurricular academic activities such as tutoring or study groups.
2. **Parental Education:**
   - Educational attainment of parents or guardians.
   - Involvement of parents in the student's academic activities and homework.
3. **Socio-Economic Factors:**
   - Family income level.
   - Access to educational resources such as books, internet, and tutoring.
   - Parent's marital status
   - How many siblings the student has
4. **Personal Factors:**
   - Student's physical and mental health.
   - Student is the first child or not
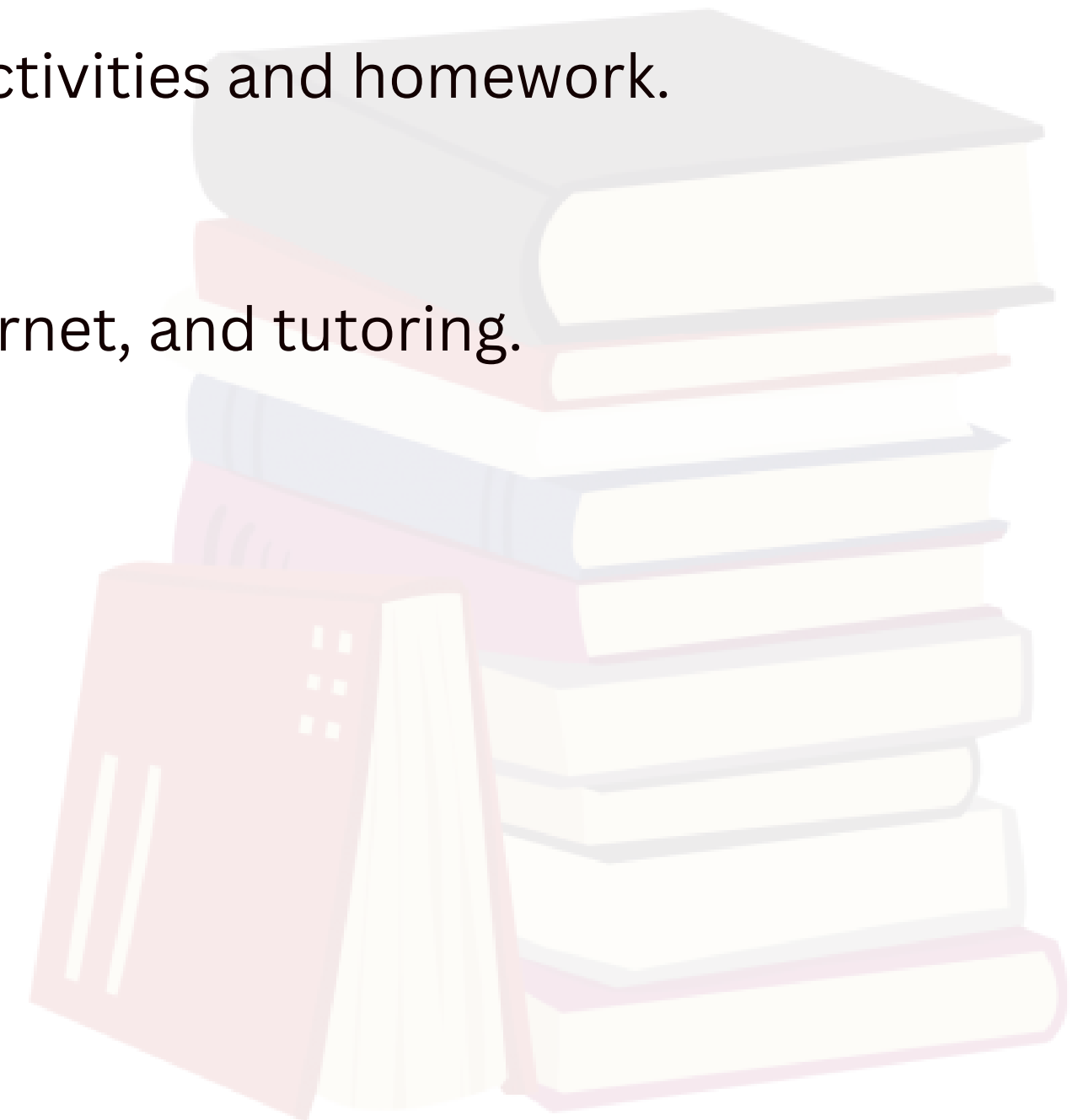
Methodology:

Data Collection and Cleaning:

Surveys and questionnaires distributed to students, parents, and teachers.
Identified the data for null or duplicate values
Removed some unwanted columns
Changed all columns in lower case better understanding
Divided the complete data into various groups to study

Data Analysis:

Statistical analysis to identify correlations and trends

Expected Outcomes:

- Identification of key factors that significantly impact student performance.
- Insights into how different variables interact and affect academic outcomes.
- Recommendations for targeted interventions to support underperforming students.
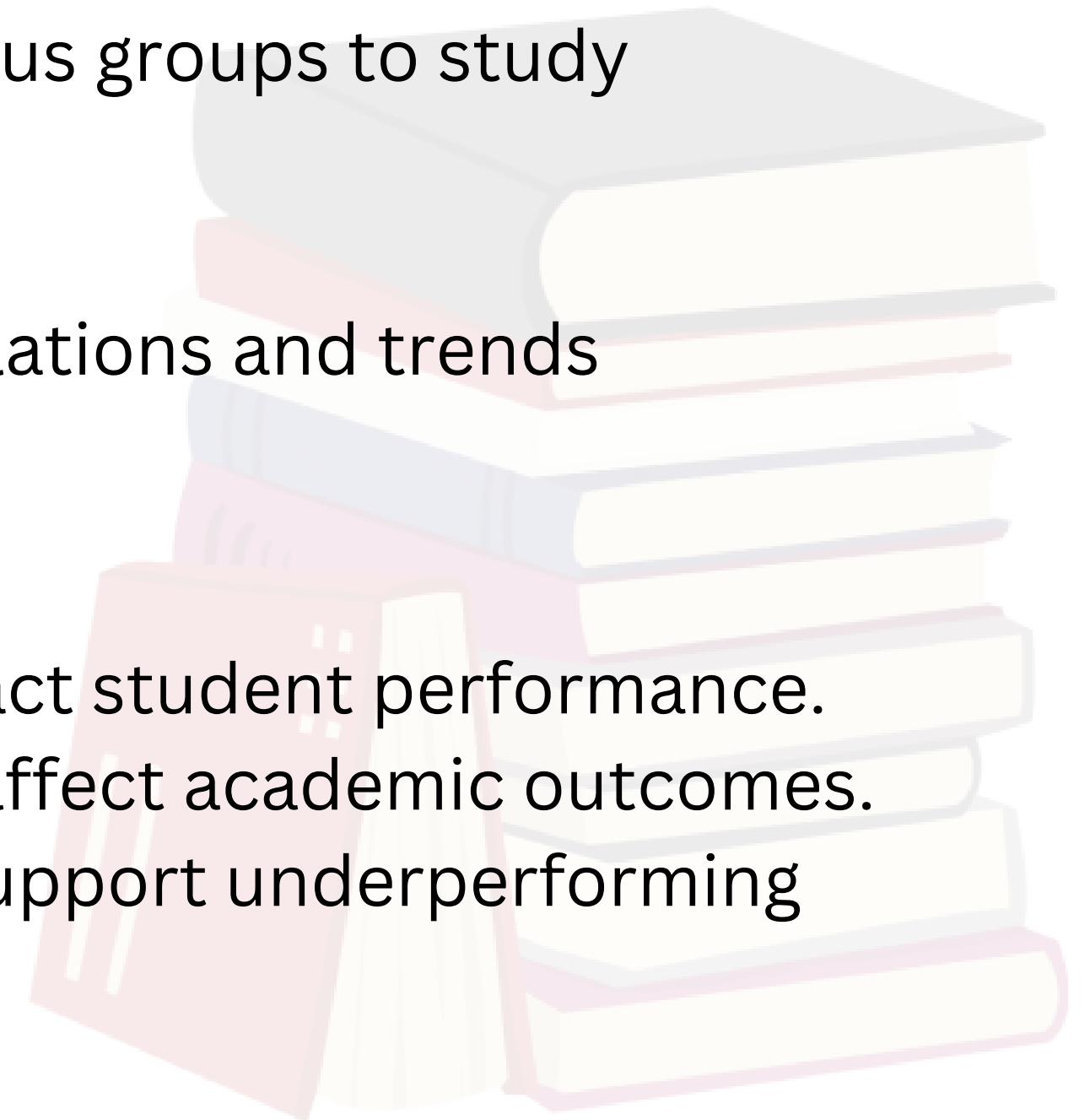
Finding Correlation between all numeric columns

```
[ ] df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
nrsiblings	29069.0	2.145894	1.458242	0.0	1.0	2.0	3.0	7.0
mathscore	30641.0	66.558402	15.361616	0.0	56.0	67.0	78.0	100.0
readingscore	30641.0	69.377533	14.758952	10.0	59.0	70.0	80.0	100.0
writingscore	30641.0	68.418622	15.443525	4.0	58.0	69.0	79.0	100.0

Here we can analyze that the maximum numbers for each subject are the same but minimum score is 0 which is scored in Math subject. Mean value is also low for Math so we will further investigate, what can be the reason of less marks in Math subject.

How many students are below the expectation level

```
[ ]  # Calculate the first quartile for each subject score
     math_Q1 = df['mathscore'].quantile(0.25)
     reading_Q1 = df['readingscore'].quantile(0.25)
     writing_Q1 = df['writingscore'].quantile(0.25)

     # Count the number of students below the first quartile for each subject
     math_below_Q1 = (df['mathscore'] < math_Q1).sum()
     reading_below_Q1 = (df['readingscore'] < reading_Q1).sum()
     writing_below_Q1 = (df['writingscore'] < writing_Q1).sum()

     # Print the number of students below the first quartile for each subject
     print("Number of students below the first quartile for each subject:")
     print("Math:", math_below_Q1)
     print("Reading:", reading_below_Q1)
     print("Writing:", writing_below_Q1)
```
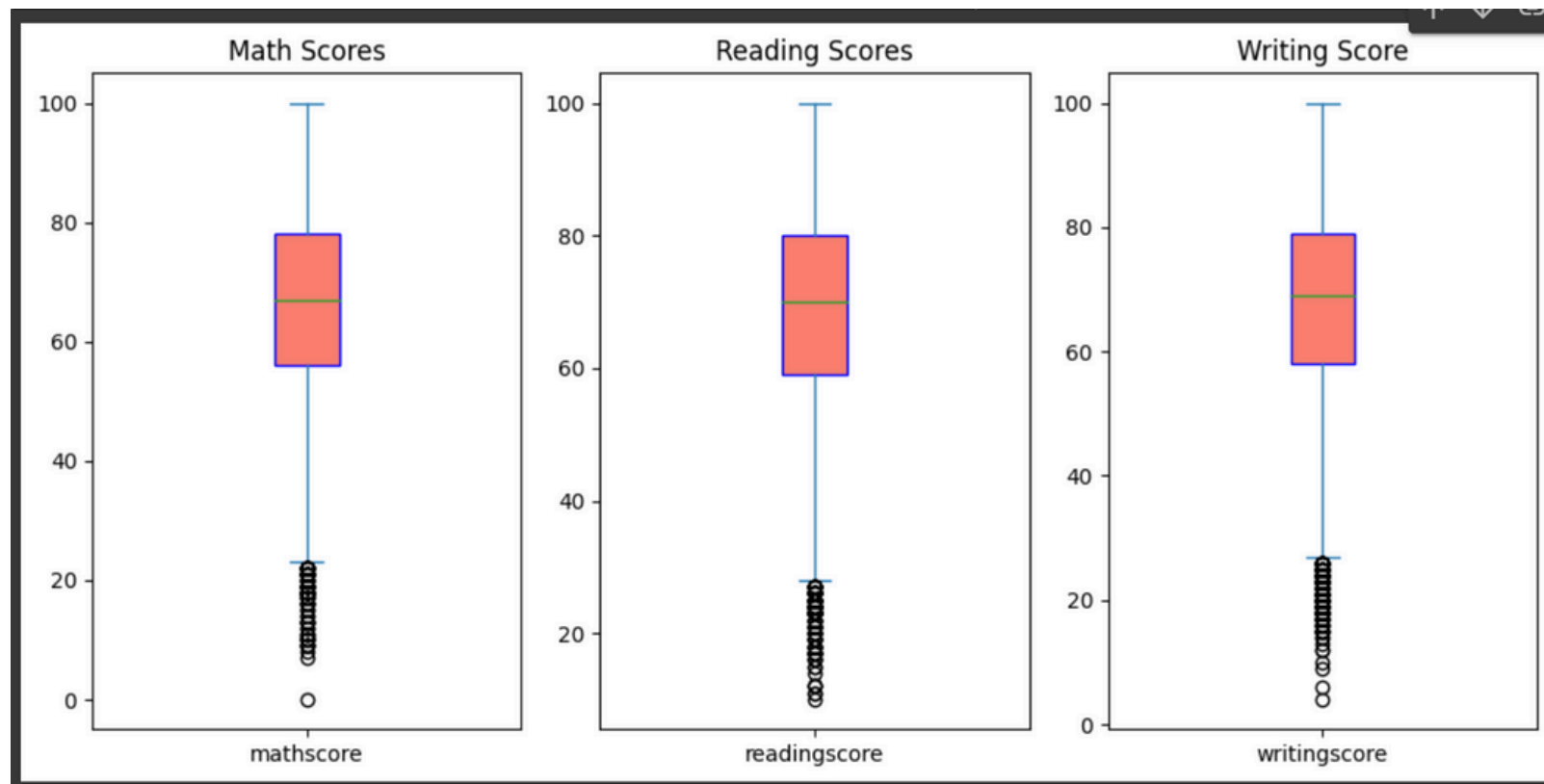
```
⇥  Number of students below the first quartile for each subject:
    Math: 7313
    Reading: 7133
    Writing: 7485
```

we have a total of 30641 students, out of them these students are below expectation level.

# How many students are below the expectation level - Outliers


```python
fig, ax = plt.subplots(ncols = 3, figsize = (10,5))
# Plot boxplots for each subject score and fill with colors
boxprops = dict(facecolor='salmon', color='blue')
score_columns['mathscore'].plot(kind='box', ax=ax[0], patch_artist=True, boxprops=boxprops)
score_columns['readingscore'].plot(kind='box', ax=ax[1], patch_artist=True, boxprops=boxprops)
score_columns['writingscore'].plot(kind='box', ax=ax[2], patch_artist=True, boxprops=boxprops)

ax[0].set_title('Math Scores')
ax[1].set_title('Reading Scores')
ax[2].set_title('Writing Score')
plt.tight_layout()
plt.show()
```

Its clear from the chart that significant number of students are scoring below what we might consider as the minimum expected level. This suggests that there is a notable portion of students who may be struggling to meet the expected standards of their academic performance. Minimum marks are 0 in Math.
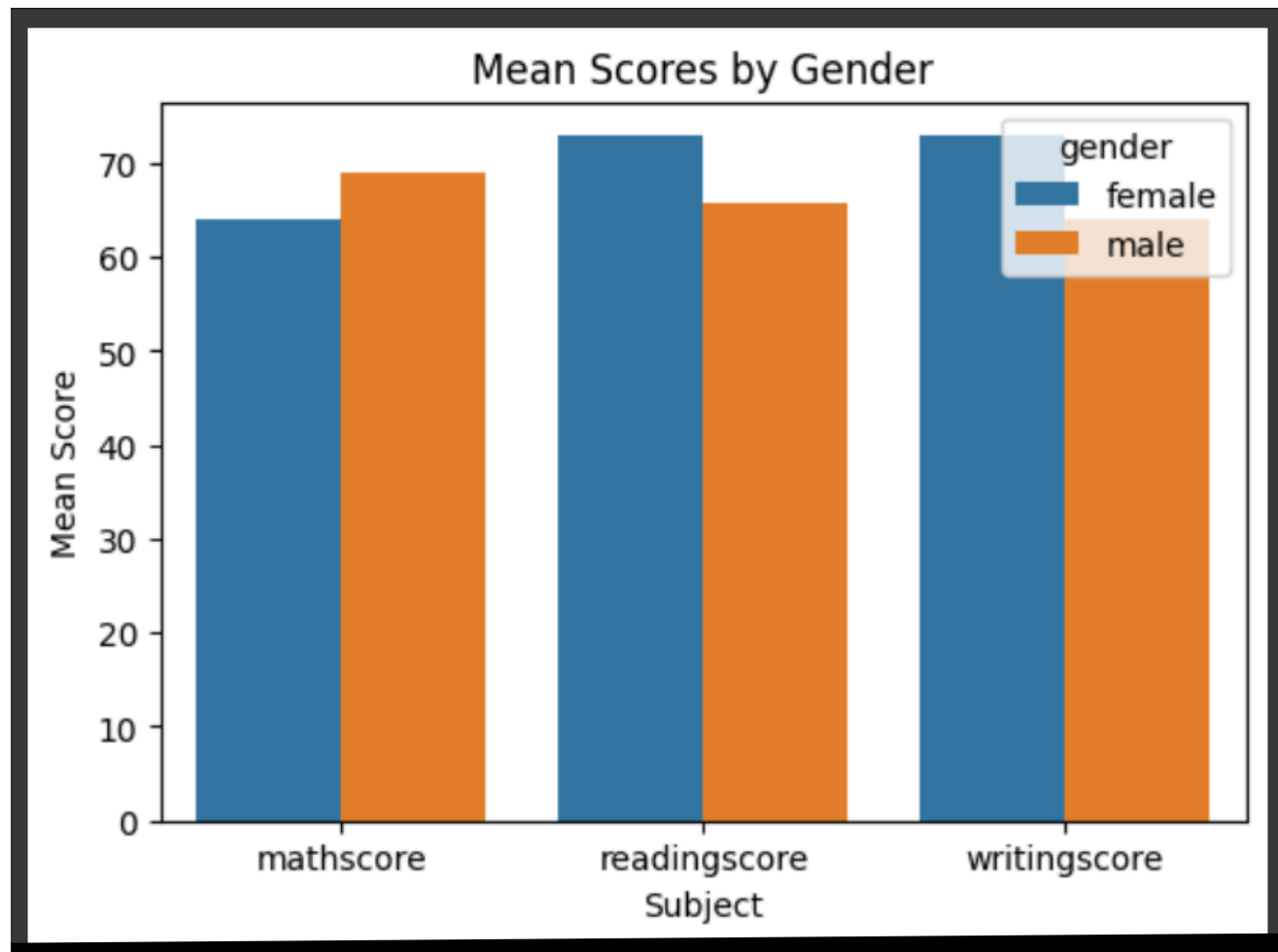
Distribution of students on basis of their group and Gender
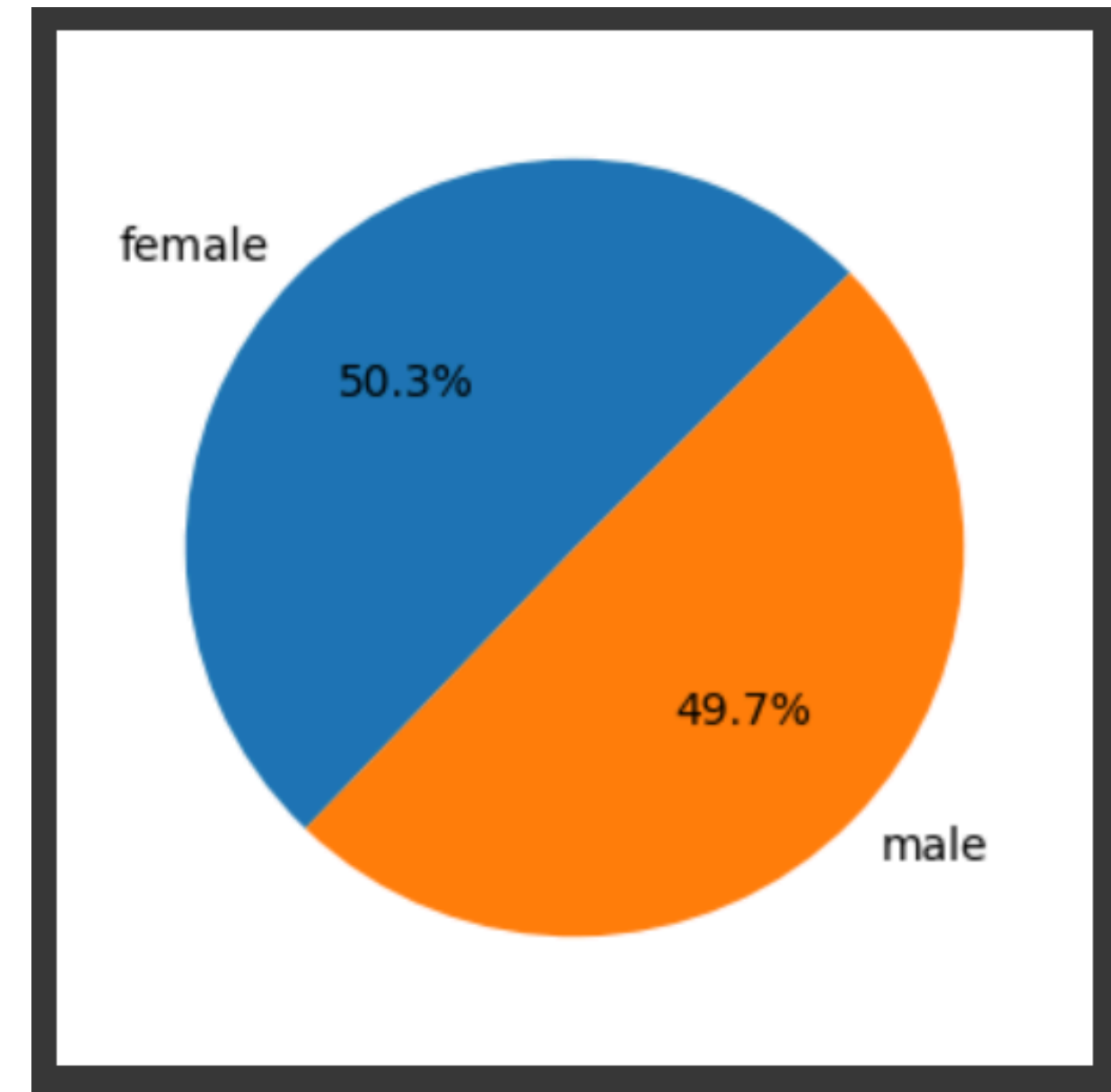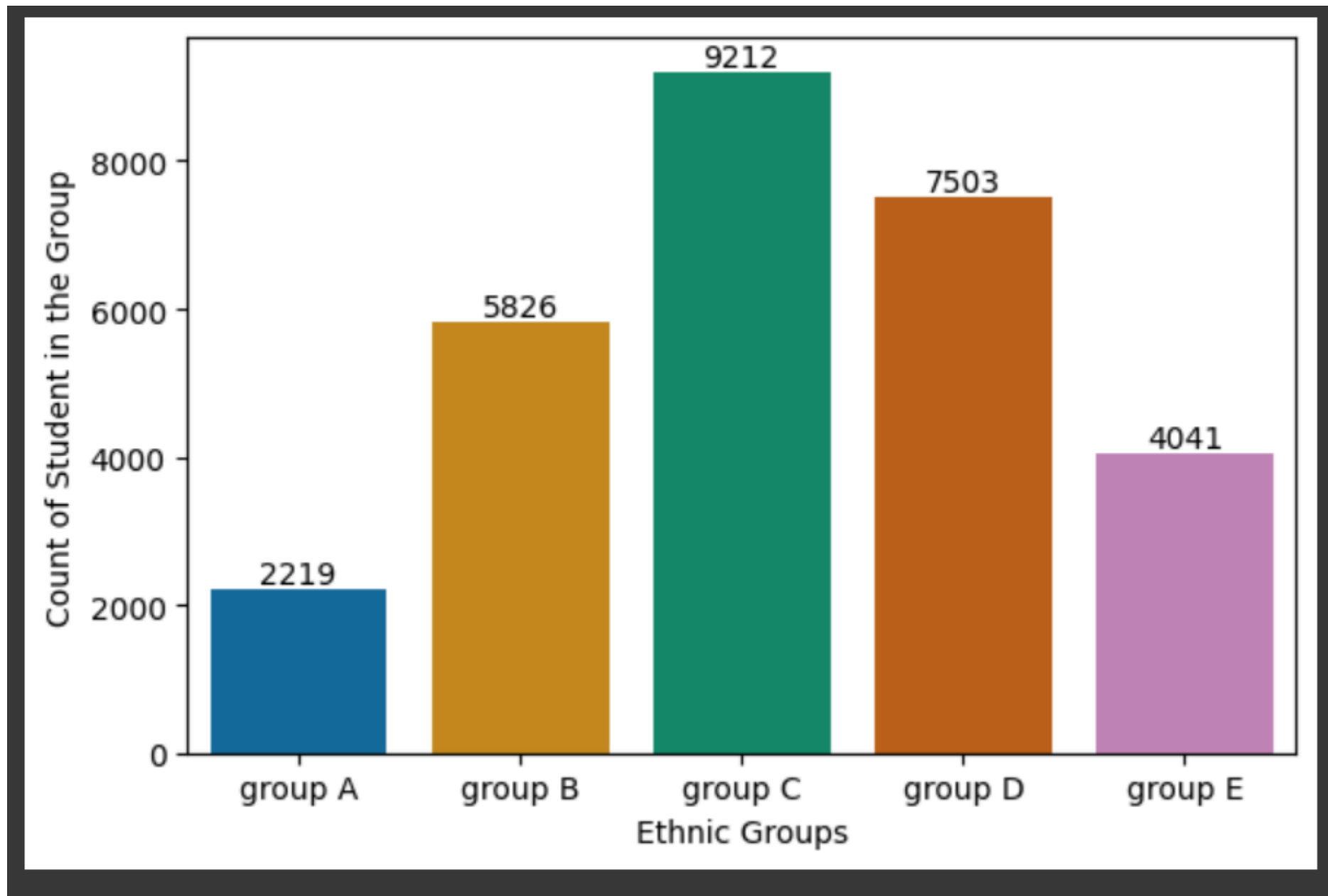
From the above chart we can analyze the distribution of students among Gender and different Groups:
- The number of female is slightly higher than males.
- Group C comprises the largest number of students whereas Group A comprises the smallest number of students.
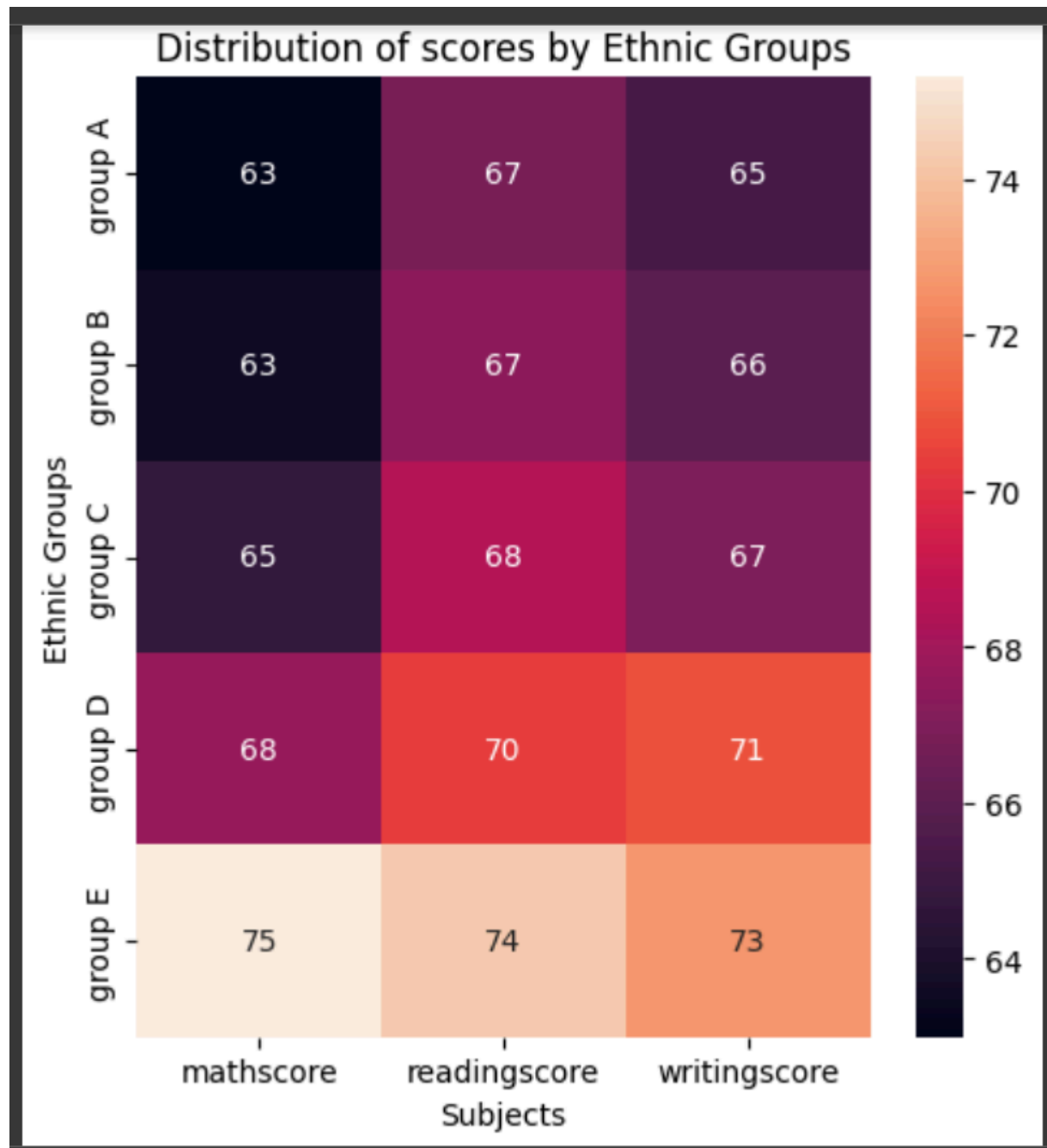
Distribution of marks on basis of Gender


```python
mean_of_Gender =
df.groupby(['gender']).agg({'mathscore':'mean','readingscore':'mean','writingscore':'mean'}).reset_index()
```

```python
# Melt the DataFrame for seaborn
mean_of_Gender_melted = mean_of_Gender.melt(id_vars='gender',
                                            var_name='subject',
                                            value_name='mean_score')

# Create the barplot
plt.figure(figsize=(6, 4))
sns.barplot(x='subject', y='mean_score', hue='gender', data=mean_of_Gender_melted)
plt.title('Mean Scores by Gender')
plt.ylabel('Mean Score')
plt.xlabel('Subject')
plt.show()
```

Upon analyzing the average marks of each subject based on gender, we found that female students excel in reading and writing, but not in mathematics.

Distribution of marks on basis of their group

Distribution of scores by Ethnic Groups

```
mean_of_ethnicgroup =
df.groupby(['ethnicgroup']).agg({'mathscore':'mean','readingscore':'mean','writingscore':'mean'})
```

```python
plt.figure(figsize = (5,5))
a = sns.heatmap(data = mean_of_ethnicgroup, annot = True)
plt.tight_layout()
plt.xlabel('Subjects')
plt.ylabel('Ethnic Groups')
plt.title('Distribution of scores by Ethnic Groups')
plt.show()
```

Upon analyzing the average scores of each subject by group, we discovered that students in Groups A and B are underperforming in mathematics.
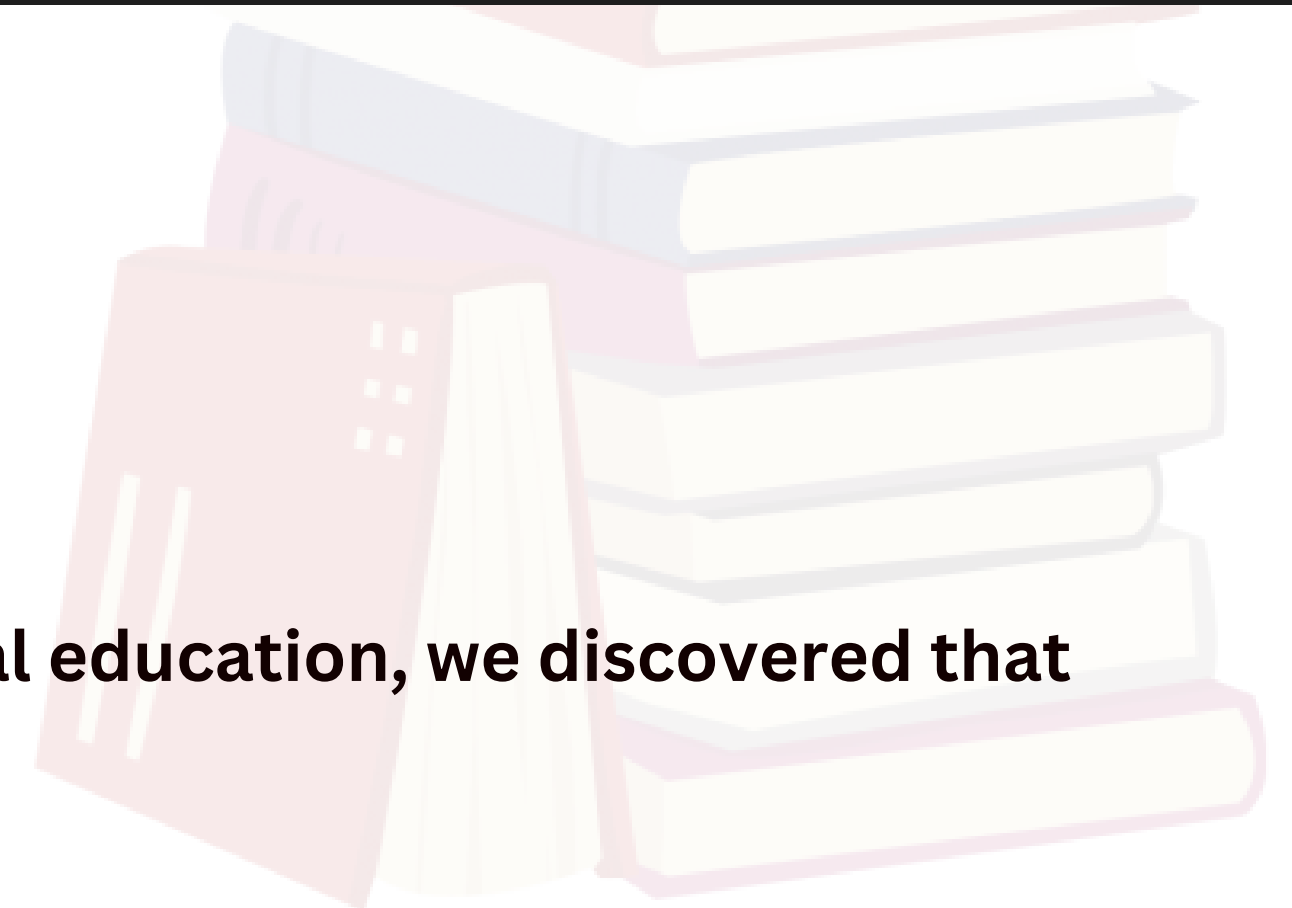
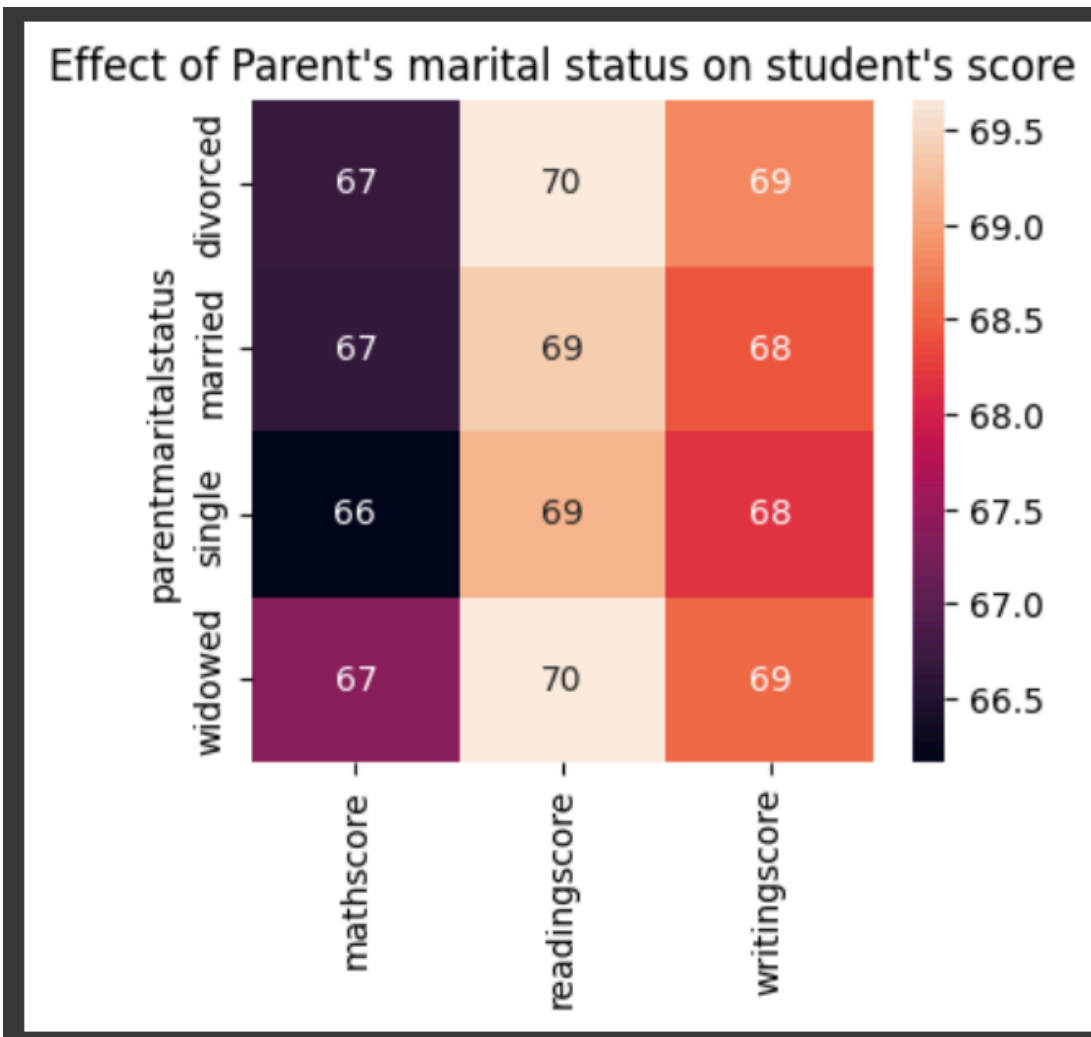# Distribution of marks on basis of their parent's education
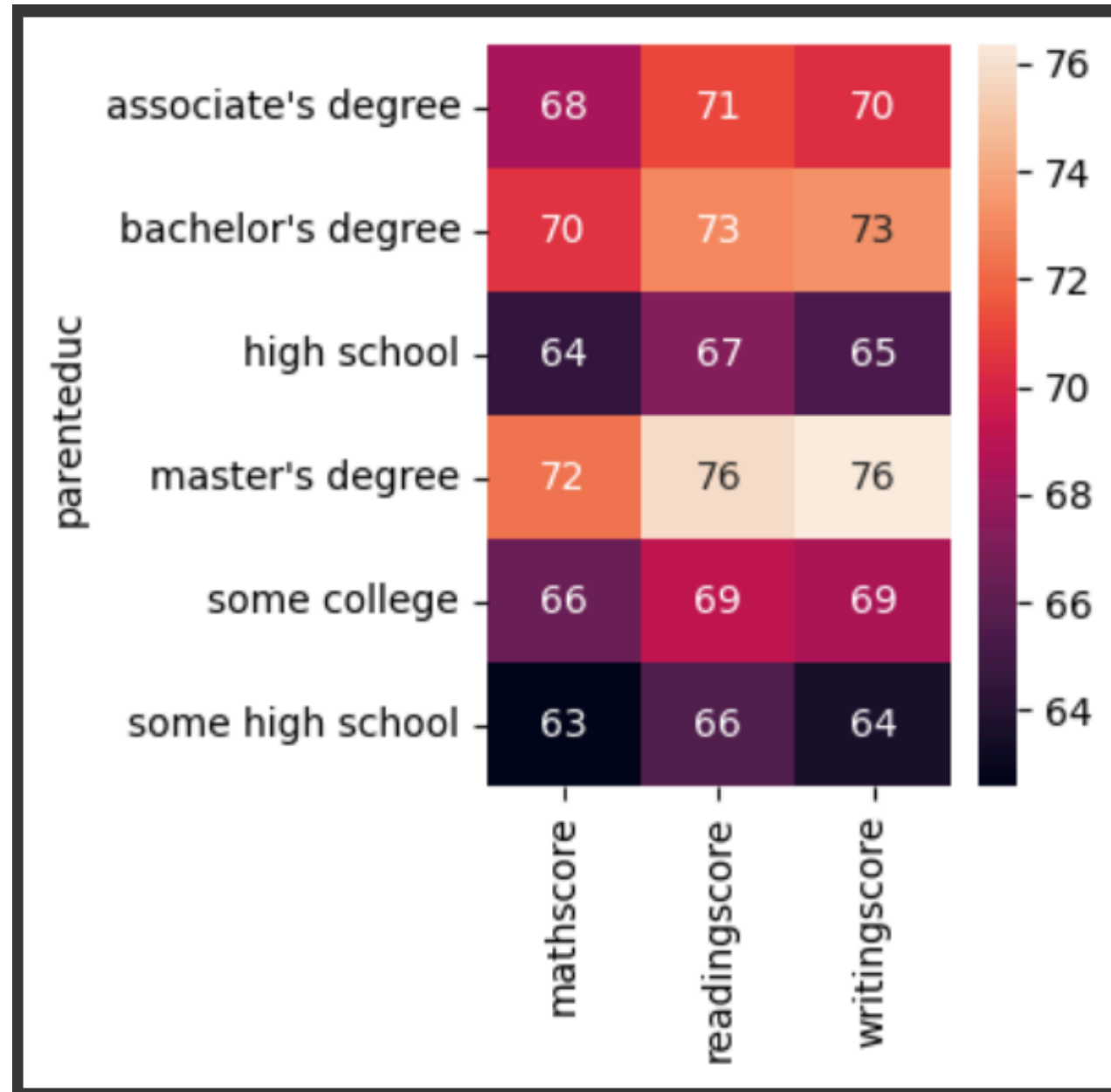

```python
gp1 =
df.groupby('parenteduc').agg({'mathscore' : 'mean', 'readingscore':'mean', 'writingscore':'mean'})
```

```python
plt.figure(figsize = (4,4))
a = sns.heatmap(data = gp1,  annot = True)
plt.tight_layout()
plt.show()
```

Upon analyzing the mean scores of each subject based on parental education, we discovered that students whose parents are well-educated achieve higher marks.
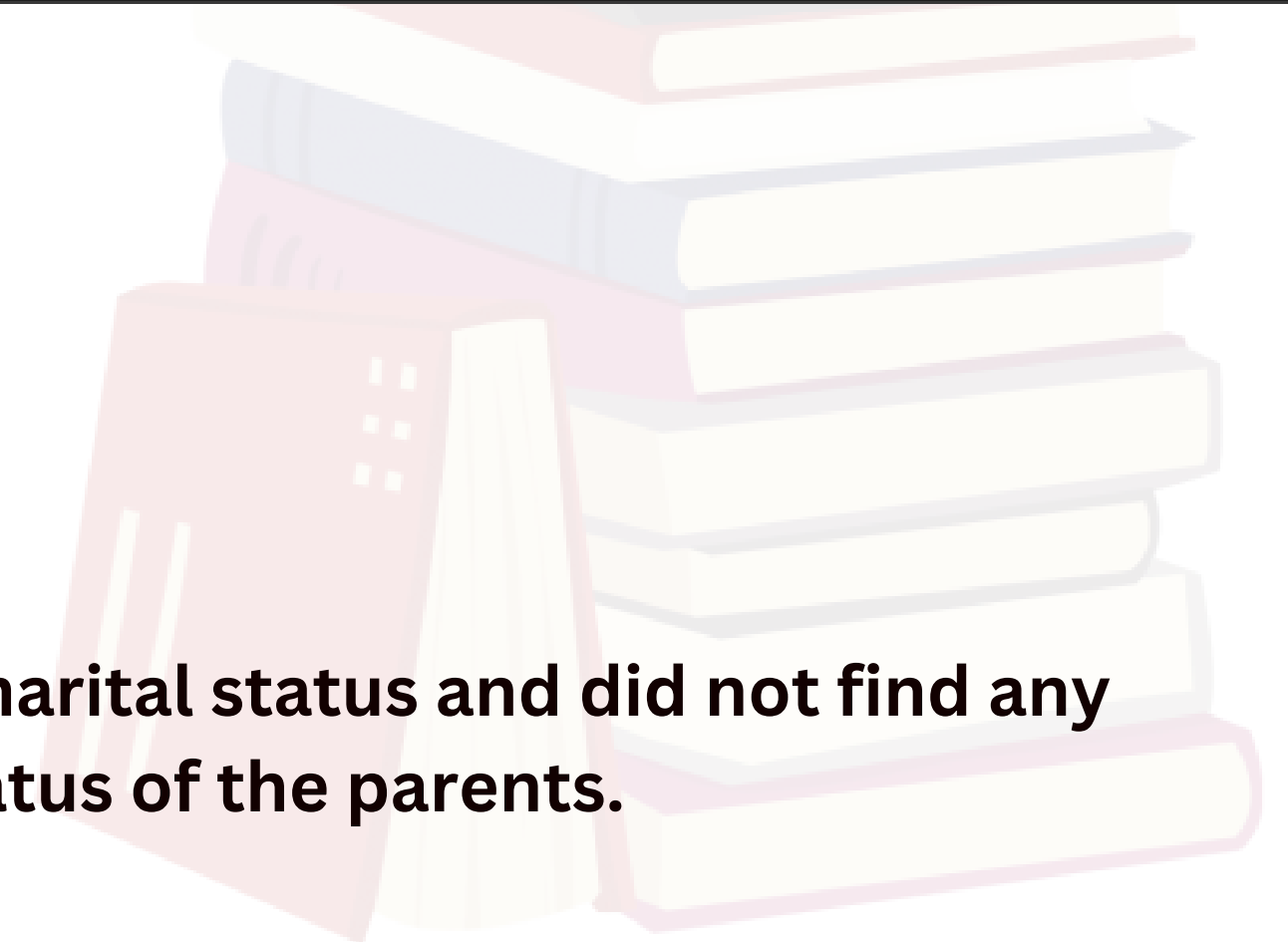
Distribution of marks on basis of their parent's marital status

Effect of Parent's marital status on student's score

```
gp2 =
df.groupby('parentmaritalstatus').agg({'mathscore' : 'mean', 'readingscore':'mean', 'writingscore':'mean'})
```

```
plt.figure(figsize = (4,4))
a = sns.heatmap(data = gp2,  annot = True)
plt.tight_layout()
plt.title('''Effect of Parent's marital status on student's score''')
plt.show()
```

We analyzed the mean scores of each subject based on parental marital status and did not find any significant differences in the scores attributable to the marital status of the parents.

Distribution of marks on basis of their participation in sports

Impact sport practice on score
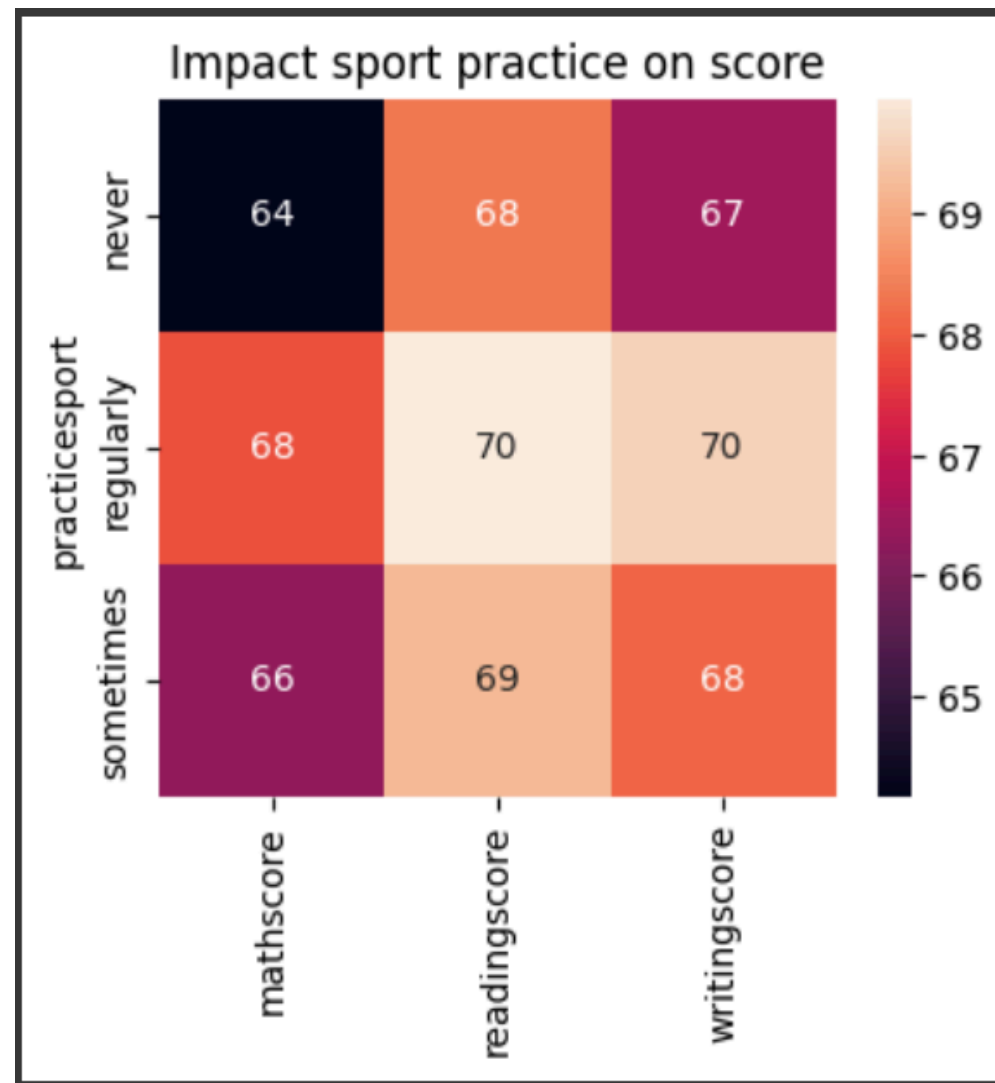
```
gp3 =
df.groupby('practicesport').agg({'mathscore' : 'mean', 'readingscore':'mean', 'writingscore':'mean'})
```

```python
# Impact of participation in sports on score
plt.figure(figsize = (4,4))
a = sns.heatmap(data = gp3, annot = True)
plt.tight_layout()
plt.title('Impact of sport practice on score')
plt.show()
```

Upon analyzing the mean scores of each subject based on sports participation, we observed that students who have never participated in sports are not proficient in mathematics.

Distribution of marks on basis of if the student is first born or not

```
gp4 = df.groupby('isfirstchild').agg({'mathscore':'mean', 'readingscore':'mean','writingscore':'mean'})
gp4
```

isfirstchild	mathscore	readingscore	writingscore
no	66.246832	69.132614	68.210887
yes	66.740646	69.542553	68.558484

Analyzing the mean scores based on whether a student is the firstborn or not, we observed no difference between the two groups. Thus, this factor does not appear to affect scores.

Distribution of marks on basis of their study hours

```
# Weekly Study Hours
gp6 = df.groupby('wklystudyhours').agg({'mathscore':'mean', 'readingscore': 'mean', 'writingscore':'mean'})
gp6
```

wklystudyhours	mathscore	readingscore	writingscore
5 - 10	66.870491	69.660532	68.636280
< 5	64.580359	68.176135	67.090192
> 10	68.696655	70.365436	69.777778

After analyzing the mean scores of each subject based on a student's weekly study hours, we found that this factor has no significant effect on students' scores.
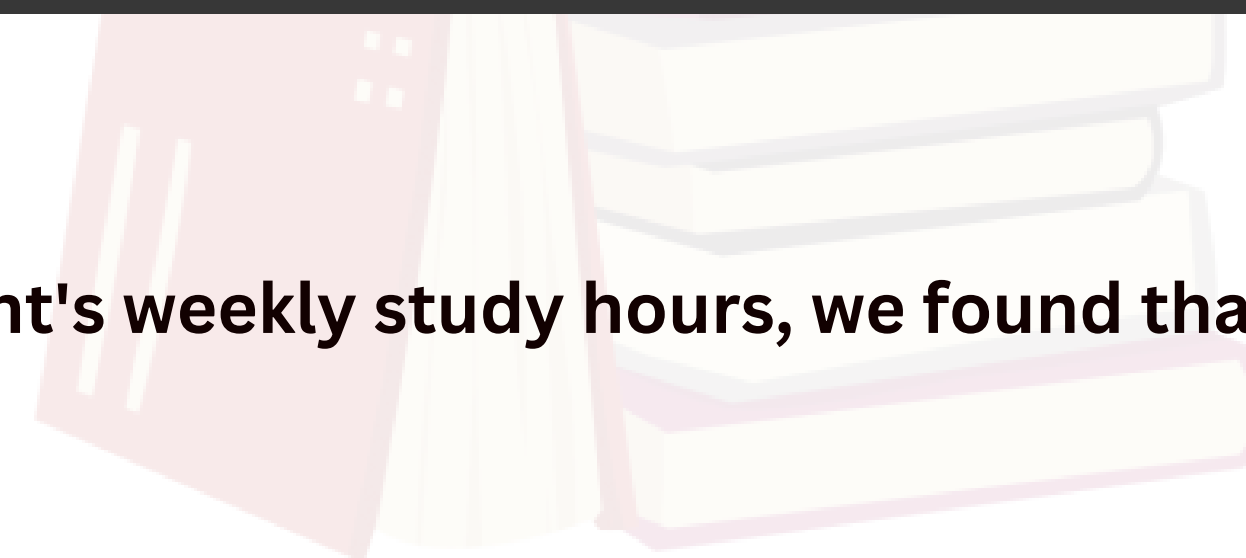
Conclusion:

Based on our comprehensive analysis of various factors affecting student performance, we have identified several key areas and potential interventions:

- **Gender and Performance in Mathematics:**
  - Our analysis shows that girls are scoring lower in mathematics. To address this, we can organize targeted study groups for girls or implement motivational programs to boost their confidence and performance in math.
- **School Groups and Math Performance:**
  - Students in Groups A and B are lagging behind in mathematics. To help these students catch up, we can provide extra classes specifically for Groups A and B.
- **Parental Education and Student Performance:**
  - There is a noticeable impact of parental education on student scores. We can organize meetings with parents to discuss how they can support their children's education more effectively.
- **Sports Participation and Academic Performance:**
  - Students who have never participated in sports are scoring lower. We should encourage these students to engage in sports or competitions, as participation in sports can enhance overall personality development
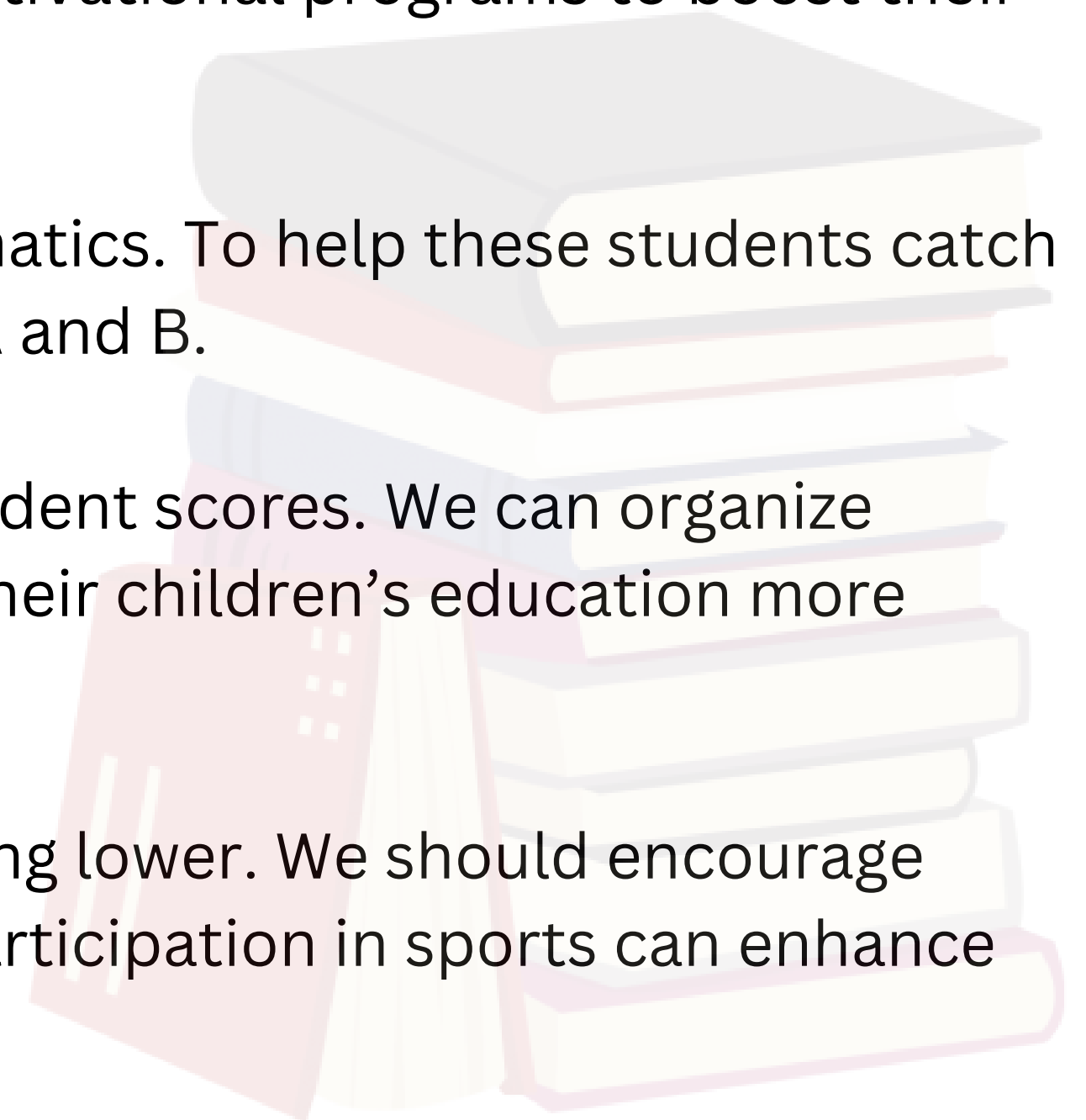
- **Factors with No Significant Impact on Scores**:
  - **Parental Marital Status and Birth Order**: Our analysis indicates that these factors do not significantly affect student scores.
  - **Weekly Study Hours**: Although weekly study hours do not significantly impact scores, we should focus on teaching students efficient study techniques to help them use their time more effectively.
- **Teacher Support and Student Guidance**:
  - Teachers should provide additional support to students who are struggling to understand subjects. As educators, they are best positioned to identify areas where individual students need improvement.
  - Additionally, we should offer counseling to all students to understand the underlying reasons for their lower scores and address those issues accordingly.

By implementing these measures, we can create a more supportive and effective learning environment that addresses the specific needs of our students and helps them achieve their full potential.