Analyzing the Impact of Demographic and Behavioral Factors on Student Academic Performance Using Python

In [21]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [22]:
```
df=pd.read_csv('/Users/deeps/Documents/projects/unguided/python/student s
```

In [4]:
```
df.shape
```

Out[4]:
```
(30641, 15)
```

In [5]:
```
df.describe(include='all')
```

Out[5]:

	Unnamed: 0	Gender	EthnicGroup	ParentEduc	LunchType	TestPrep	Pa
count	30641.000000	30641	28801	28796	30641	28811	
unique	NaN	2	5	6	2	2	
top	NaN	female	group C	some college	standard	none	
freq	NaN	15424	9212	6633	19905	18856	
mean	499.556607	NaN	NaN	NaN	NaN	NaN	
std	288.747894	NaN	NaN	NaN	NaN	NaN	
min	0.000000	NaN	NaN	NaN	NaN	NaN	
25%	249.000000	NaN	NaN	NaN	NaN	NaN	
50%	500.000000	NaN	NaN	NaN	NaN	NaN	
75%	750.000000	NaN	NaN	NaN	NaN	NaN	
max	999.000000	NaN	NaN	NaN	NaN	NaN	

In [6]:
```
df.head(3)
```

Out[6]:

	Unnamed: 0	Gender	EthnicGroup	ParentEduc	LunchType	TestPrep	ParentMarit
0	0	female	NaN	bachelor's degree	standard	none	
1	1	female	group C	some college	standard	NaN	
2	2	female	group B	master's degree	standard	none	

In [7]:
```
#  Dropping unnamed column as it is same as index column

df.drop('Unnamed: 0',axis=1,inplace=True)
df.head(2)
```

Out[7]:

	Gender	EthnicGroup	ParentEduc	LunchType	TestPrep	ParentMaritalStatus	P
0	female	NaN	bachelor's degree	standard	none	married	
1	female	group C	some college	standard	NaN	married	

In [8]:
```
df.dtypes
```

Out[8]:
```
Gender                 object
EthnicGroup            object
ParentEduc             object
LunchType              object
TestPrep               object
ParentMaritalStatus    object
PracticeSport          object
IsFirstChild           object
NrSiblings            float64
TransportMeans         object
WklyStudyHours         object
MathScore               int64
ReadingScore            int64
WritingScore            int64
dtype: object
```

Checking for null values

In [9]:
```
df.isnull().sum()
```

Out[9]:
```
Gender                    0
EthnicGroup            1840
ParentEduc             1845
LunchType                 0
TestPrep               1830
ParentMaritalStatus    1190
PracticeSport           631
IsFirstChild            904
NrSiblings             1572
TransportMeans         3134
WklyStudyHours          955
MathScore                 0
ReadingScore              0
WritingScore              0
dtype: int64
```

Score columns that are required for result analysis have no null values , so null values in all other columns is considered as missing data

In [10]:
```
df.duplicated().sum()
```

Out[10]:
```
np.int64(0)
```

In [ ]:
```
df.shape
```

Out[ ]:
```
(30641, 14)
```

1. Average Scores By Gender

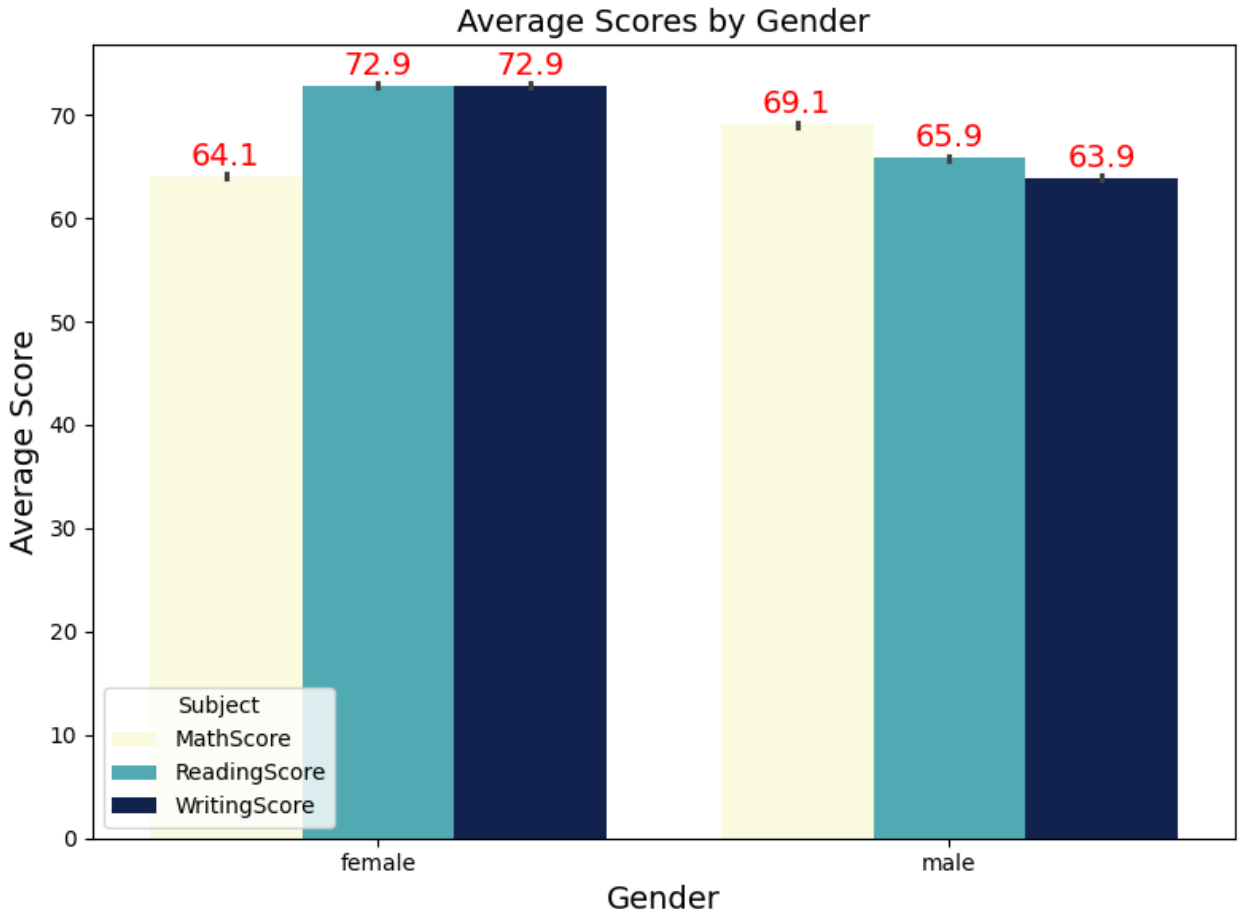
In [85]:
```
scores_gender = pd.melt(df, id_vars='Gender',
                        value_vars=['MathScore', 'ReadingScore', 'WritingScore']
                        var_name='Subject', value_name='Score')
colors = list(plt.cm.YlGnBu(np.linspace(0, 1, 3)))


plt.figure(figsize=(8, 6))
bp = sns.barplot(x='Gender', y='Score', hue='Subject', data=scores_gender

plt.title("Average Scores by Gender", fontsize=14)
plt.xlabel("Gender", fontsize=14)
plt.ylabel("Average Score", fontsize=14)
plt.legend(title="Subject")

for container in bp.containers:
    bp.bar_label(container, fmt='%.1f', label_type='edge', fontsize=14, p

plt.tight_layout()
plt.show()
```



Insights:

- **Math**: Male students have a higher average (69.1) compared to females (64.1).
- **Reading**: Female students significantly outperform males, scoring 72.9 on average >compared to 65.9.
- **Writing**: Female students again lead with an average score of 72.9, while males score 63.9.
- **Males perform slightly better in Math**.
- **Females consistently outperform males in Reading and Writing**.

Recommendations:

Support for Male Students in Literacy:

- Provide **targeted programs** to improve reading and writing skills.
- Encourage reading habits and creative writing through clubs or assignments tailored to their interests.

Balanced Curriculum:

> - Design activities that foster both logical (math) and verbal (language) skills across genders.

2.Average Scores By Ethnic Group

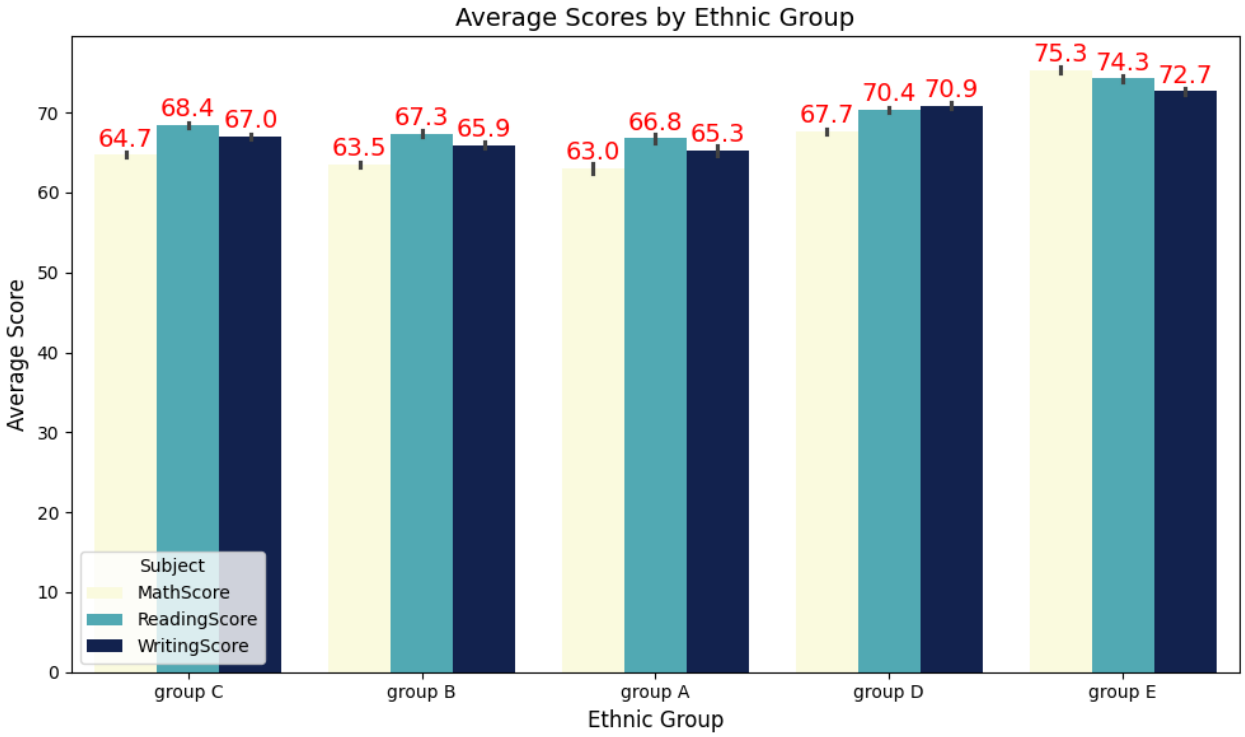
In [63]:
```
scores_ethnic = pd.melt(df, id_vars='EthnicGroup',
                        value_vars=['MathScore', 'ReadingScore', 'Writing
                        var_name='Subject', value_name='Score')
colors = list(plt.cm.YlGnBu(np.linspace(0, 1, 3)))

plt.figure(figsize=(10, 6))
eg = sns.barplot(x='EthnicGroup', y='Score', hue='Subject', data=scores_e

plt.title("Average Scores by Ethnic Group", fontsize=14)
plt.xlabel("Ethnic Group", fontsize=12)
plt.ylabel("Average Score", fontsize=12)
plt.legend(title="Subject")

for container in eg.containers:
    eg.bar_label(container, fmt='%.1f', label_type='edge', fontsize=14, p

plt.tight_layout()
plt.show()
```



> This bar chart shows the average Math, Reading, and Writing scores across five ethnic groups (A to E).
>
> ### Insights:
>
> - **Group E** has the **highest average scores** across all three subjects:
> - Math: 75.3, Reading: 74.3, Writing: 72.7

- **Group A** has the **lowest average scores**:
- Math: 63.0, Reading: 66.8, Writing: 65.3
- **Group D** also performs well, with average scores in the 67–71 range.
- Groups **B** and **C** fall in the middle, with relatively balanced scores.

Recommendations:

- **Targeted Support for Lower-Performing Groups**:
- Groups A and B may benefit from additional academic support such as tutoring, mentorship, or after-school programs.
- **Leverage High-Performing Group Strategies**:
- Study teaching methods, learning environments, or parental involvement practices in groups D and E to replicate success in other groups.
- **Culturally Responsive Education**:
- Develop instructional materials that are inclusive and relevant to all ethnic backgrounds to improve engagement and performance.

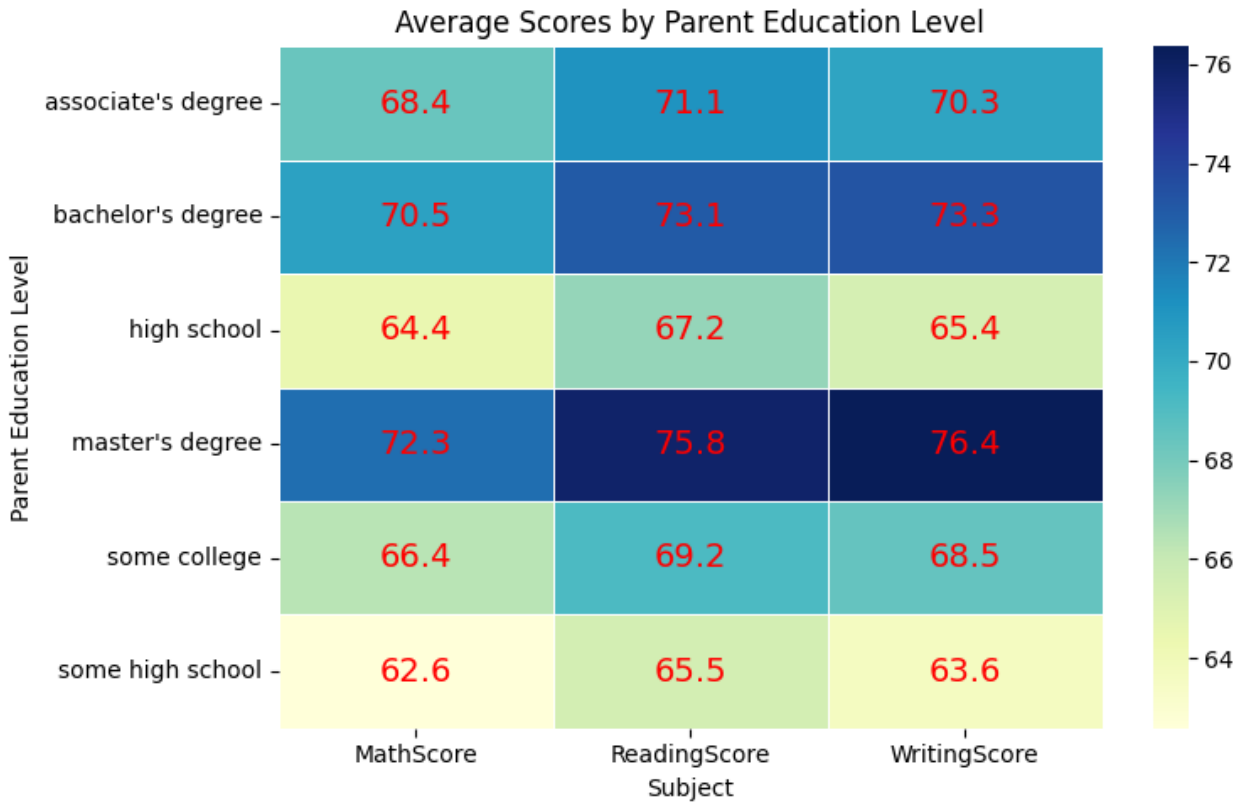
Understanding how performance varies across demographic groups allows educators to design fair and effective support strategies for all students.

3.Average Student Sores by Parent Education Level

In [66]:
```
heatmap_data = df.groupby('ParentEduc')[['MathScore', 'ReadingScore', 'Wr

plt.figure(figsize=(8, 5))
sns.heatmap(heatmap_data, annot=True, cmap='YlGnBu', fmt=".1f", linewidth

plt.title("Average Scores by Parent Education Level")
plt.ylabel("Parent Education Level")
plt.xlabel("Subject")
plt.tight_layout()
plt.show()
```



> The heatmap above illustrates the average student scores across Math, Reading, and Writing, segmented by the highest education level attained by their parents.
>
> ### Key Insights:
>
> - Students whose parents have a **master's degree** achieved the **highest average scores** in all three subjects Math (72.3), Reading (75.8), and Writing (76.4).
> - A clear **positive correlation** exists between parent education level and student /performance. As the education level increases, student scores tend to rise.
> - Students with parents who have **some high school education** show the **lowest scores** across all subjects.
> - The effect is **most pronounced in reading and writing**, suggesting that parental education may have a stronger influence on literacy development than on math.
>
> ### Recommendations:
>
> - **Parental Engagement Programs**: Schools should develop programs to help parents support their children's education, regardless of their academic background.
> - **Early Intervention**: Identify and provide extra support for students whose parents have lower education levels.
> - **Community Learning Support**: Encourage family literacy programs and workshops in communities with lower average

parental education.

- **Customized Learning Plans**: Teachers can use this data to personalize instruction, allocating more resources or attention to students who might be at an educational disadvantage.

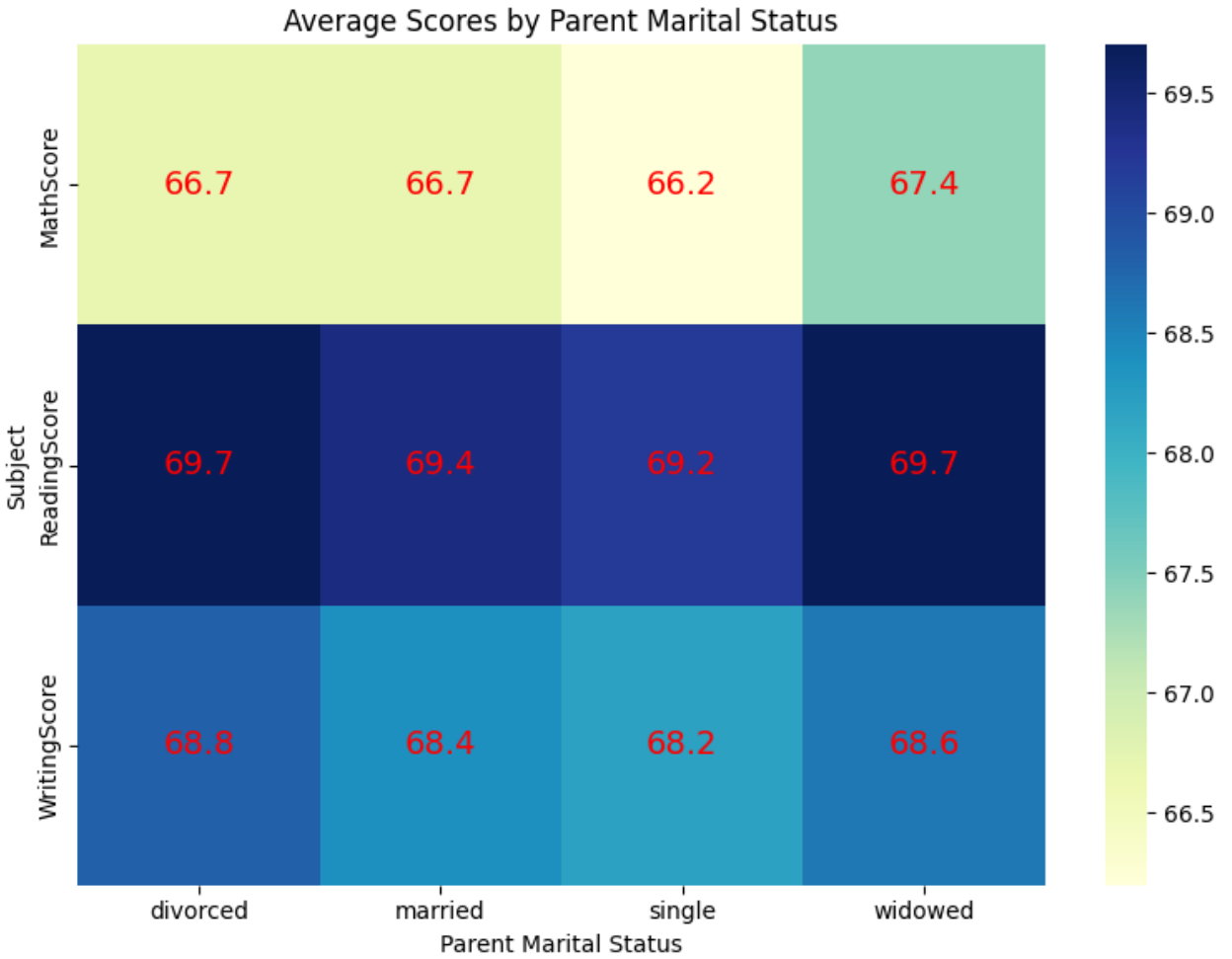
This analysis highlights how socioeconomic and educational factors in the home environment can significantly influence academic outcomes.

4.Impact of parent marital status on scores

In [67]:
```
heatmap_data = df.groupby('ParentMaritalStatus')[['MathScore', 'ReadingSc

heatmap_data = heatmap_data.T

plt.figure(figsize=(8, 6))
sns.heatmap(heatmap_data, annot=True, cmap='YlGnBu', fmt=".1f",annot_kws=
plt.title("Average Scores by Parent Marital Status")
plt.xlabel("Parent Marital Status")
plt.ylabel("Subject")
plt.tight_layout()
plt.show()
```



Insights: Impact of Parent Marital Status on Student Performance

- **Reading scores** remain consistently high across all marital statuses, with divorced and widowed parent's children slightly outperforming others (69.7 average).
- **Writing scores** also show minor variation, with children of divorced parents scoring the highest (68.8), and children of single parents slightly lower (68.2).
- **Math scores** are more variable, with the lowest among single parents (66.2) and the highest for widowed parents (67.4).

Recommendations:

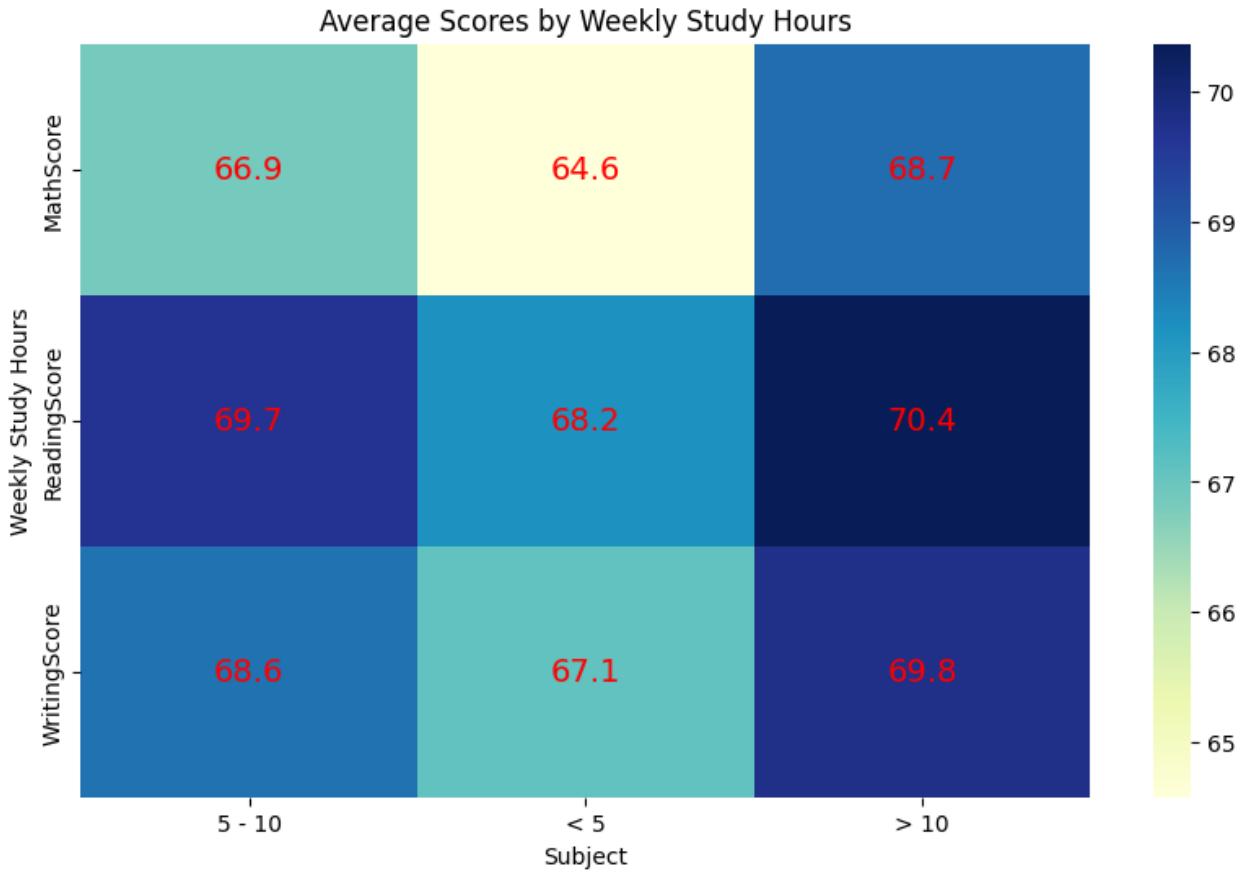
- Provide **extra academic support** in math for students from single-parent households to close performance gaps.
- Since reading and writing scores are relatively stable, focus **interventions primarily on math enrichment**.
- Consider **offering parent-focused engagement programs** to strengthen educational support at home, especially where only one parent is present.

5. Weekly Study Hours vs. Academic Performance

In [68]:
```
study_hours_scores = df.groupby("WklyStudyHours")[["MathScore", "ReadingS

study_hours_scores = study_hours_scores.T

plt.figure(figsize=(10, 6))
sns.heatmap(study_hours_scores, annot=True, cmap="YlGnBu", fmt=".1f",anno
plt.title("Average Scores by Weekly Study Hours")
plt.xlabel("Subject")
plt.ylabel("Weekly Study Hours")
plt.show()
```

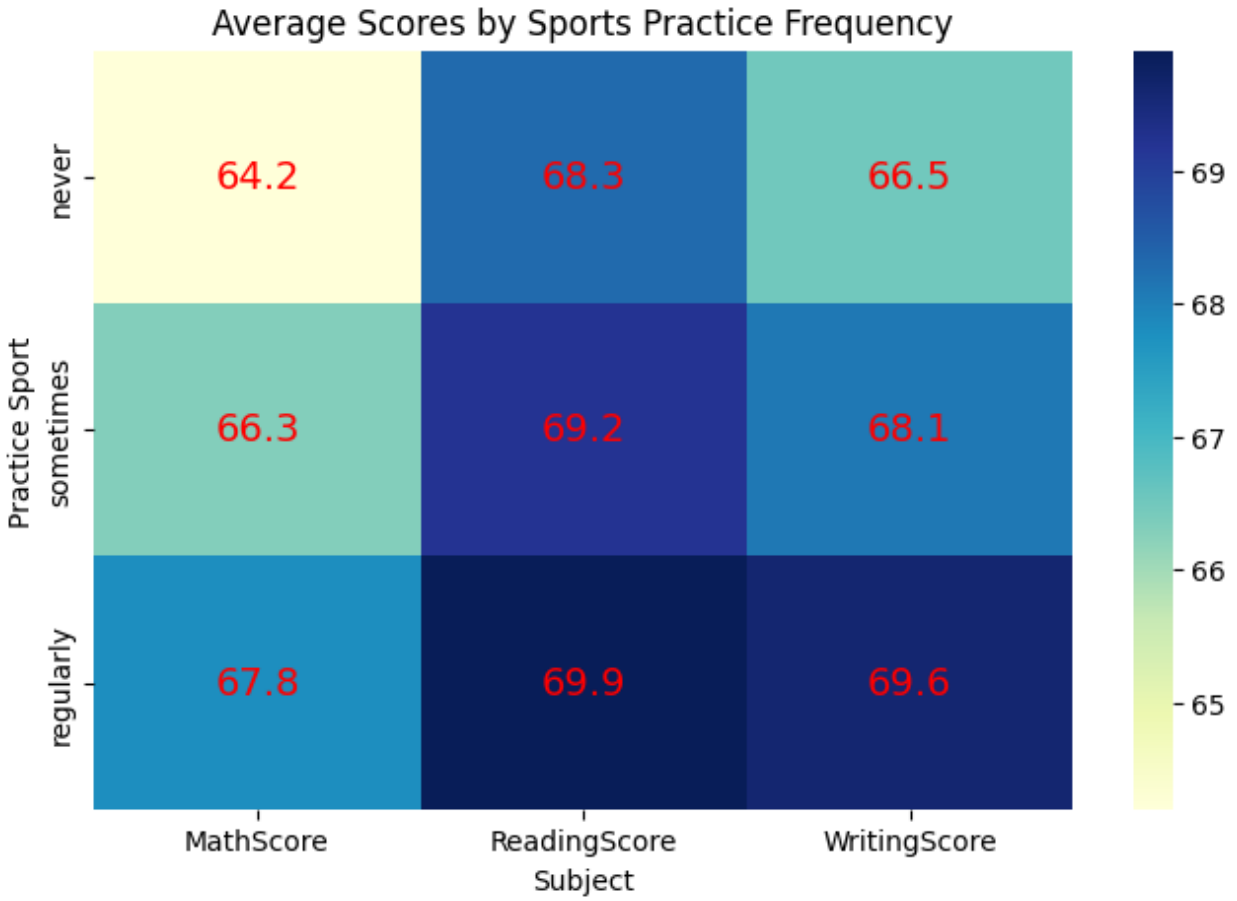


> The heatmap illustrates the relationship between students' weekly study hours and their average scores in Math, Reading, and Writing.
>
> ### Interpretation:
>
> - Students who study **more than 10 hours per week** tend to have the **highest average scores** across all subjects.
> - Math: 68.7
> - Reading: 70.4
> - Writing: 69.8
> - Those studying **less than 5 hours** have the **lowest scores**, particularly in Math (64.6).
> - **5–10 hours/week** is associated with moderate performance, slightly better than <5 hours but lower than >10.
>
> ### Recommendations:
>
> - **Encourage longer study durations**: Promote a habit of studying more than 10 hours weekly to maximize academic performance.
> - **Target underperforming groups**: Offer structured support or time management workshops for students studying less than 5 hours per week.
> - **Balance is key**: Ensure that increased study hours are accompanied by quality and not just quantity.

6.Impact of sports on student scores

In [69]:
```
pivot = df.groupby("PracticeSport")[["MathScore", "ReadingScore", "Writin
pivot = pivot.reindex(["never", "sometimes", "regularly"])

plt.figure(figsize=(8, 5))
sns.heatmap(pivot, annot=True, cmap="YlGnBu", fmt=".1f",annot_kws={"size"
plt.title("Average Scores by Sports Practice Frequency")
plt.xlabel("Subject")
plt.ylabel("Practice Sport")
plt.show()
```



Insights

- Regular sports practice is associated with **higher scores** in all subjects.
- **Math scores improve** the most (↑3.6 points from 'never' to 'regularly').
- Gains in **reading and writing** are also evident with more frequent activity.

Recommendations

- Encourage **daily or weekly sports** programs in schools.
- Highlight the **academic benefits of physical activity** to students

> and parents.
> - Promote **balance** between physical wellness and academic achievement.

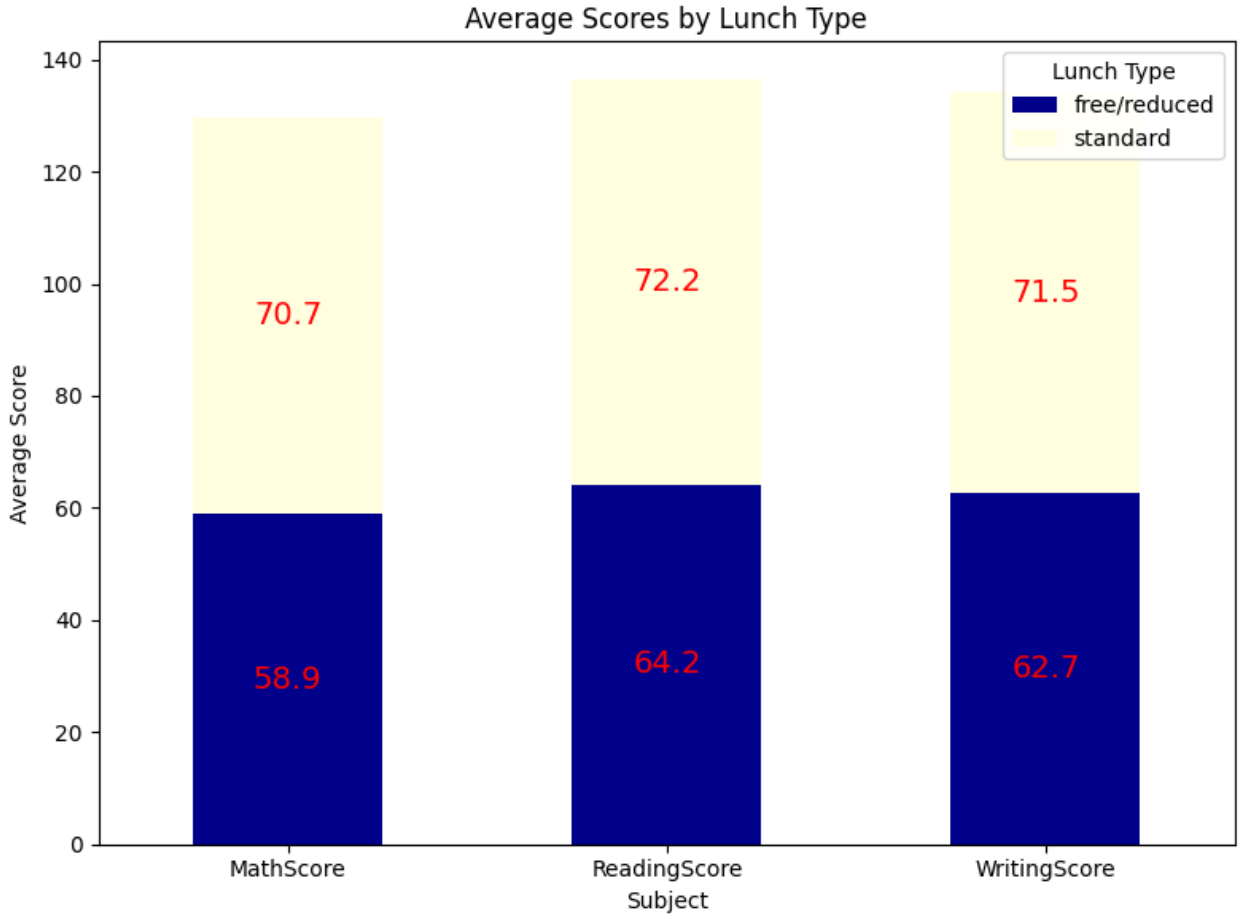
7. Average Student Scores By Lunch Type

In [73]:
```
avg_scores_lunch = df.groupby('LunchType')[['MathScore', 'ReadingScore',
avg_scores_lunch_T = avg_scores_lunch.T

lt = avg_scores_lunch_T.plot(kind='bar', stacked=True, figsize=(8, 6), co
plt.title('Average Scores by Lunch Type')
plt.ylabel('Average Score')
plt.xticks(rotation=0)
plt.xlabel('Subject')
plt.legend(title='Lunch Type')

for container in lt.containers:
    lt.bar_label(container, label_type='center', fontsize=14,color='red')

plt.tight_layout()
plt.show()
```



> ## Insights
>
> - Students receiving a **standard lunch** outperform those on **free/reduced lunch** by a significant margin.

- The largest score gap appears in **math** (an 11.8-point difference).
- Differences may reflect broader socioeconomic disparities influencing nutrition, learning environments, or access to resources.

Recommendations

- Schools should provide **nutritionally balanced meals** to all students, regardless of economic status.
- Consider **after-school academic support** for students on free/reduced lunch programs.
- Implement programs that **address socioeconomic barriers** to academic success, including breakfast provision and tutoring.

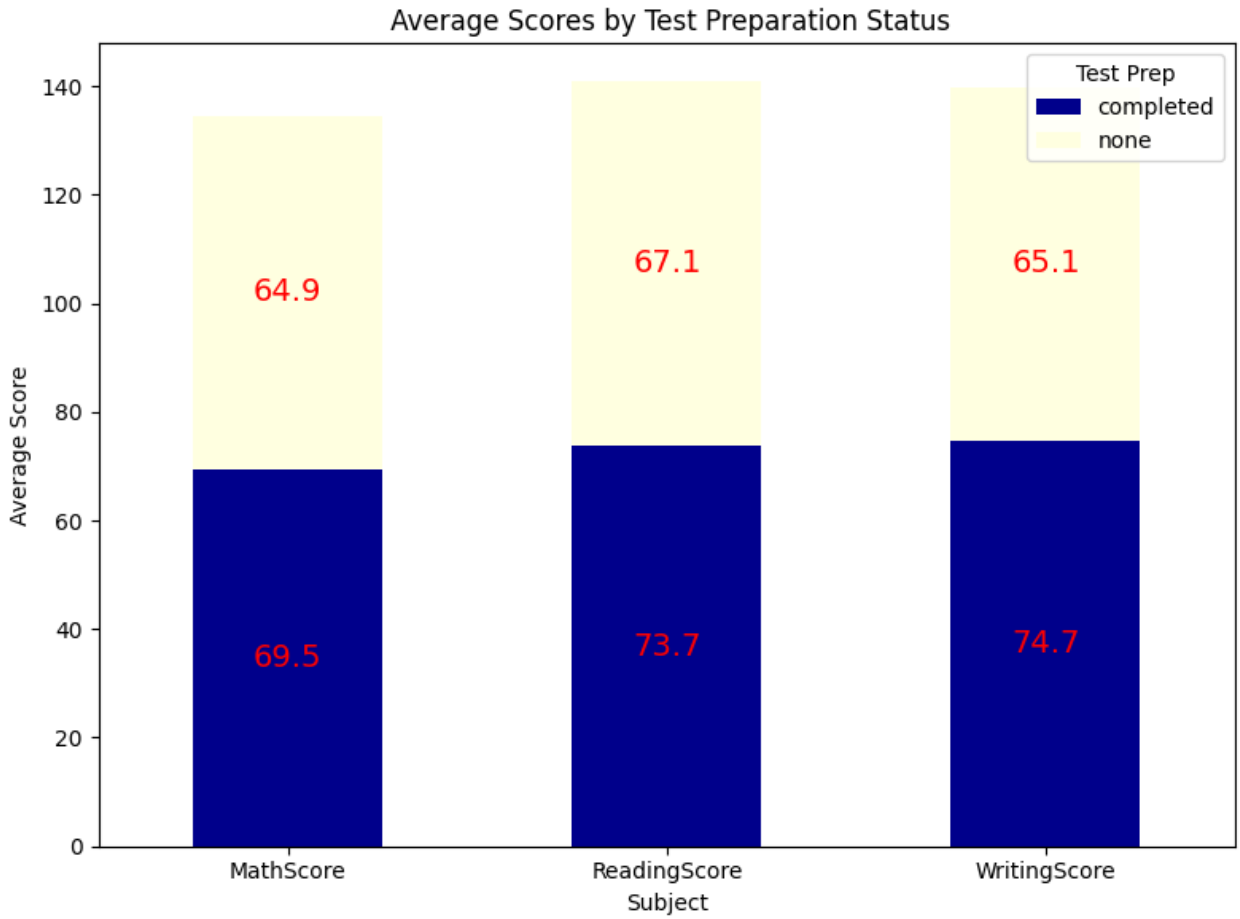
8.Average Scores By Test Preparation Status

In [89]:
```
avg_scores_prep = df.groupby('TestPrep')[['MathScore', 'ReadingScore', 'W
avg_scores_prep_T = avg_scores_prep.T

tp = avg_scores_prep_T.plot(kind='bar', stacked=True, figsize=(8, 6), col
plt.title(' Average Scores by Test Preparation Status')
plt.ylabel('Average Score')
plt.xlabel('Subject')
plt.xticks(rotation=0)
plt.legend(title='Test Prep')

for container in tp.containers:
    tp.bar_label(container, label_type='center', fontsize=14,color='red')

plt.tight_layout()
plt.show()
```



Insights

- Students who **completed the test preparation course** performed significantly better in all subjects.
- The greatest improvement is seen in **writing** scores (↑9.6 points).
- This suggests test prep courses may enhance **test-taking strategies, confidence, and content mastery**.

Recommendations

- Encourage wider **enrollment in test prep programs**, especially for underperforming students.
- Consider offering **free or subsidized test prep resources** for students with financial constraints.
- Integrate **test preparation modules** into the regular curriculum to boost readiness.

9.Average Scores By First Child Status

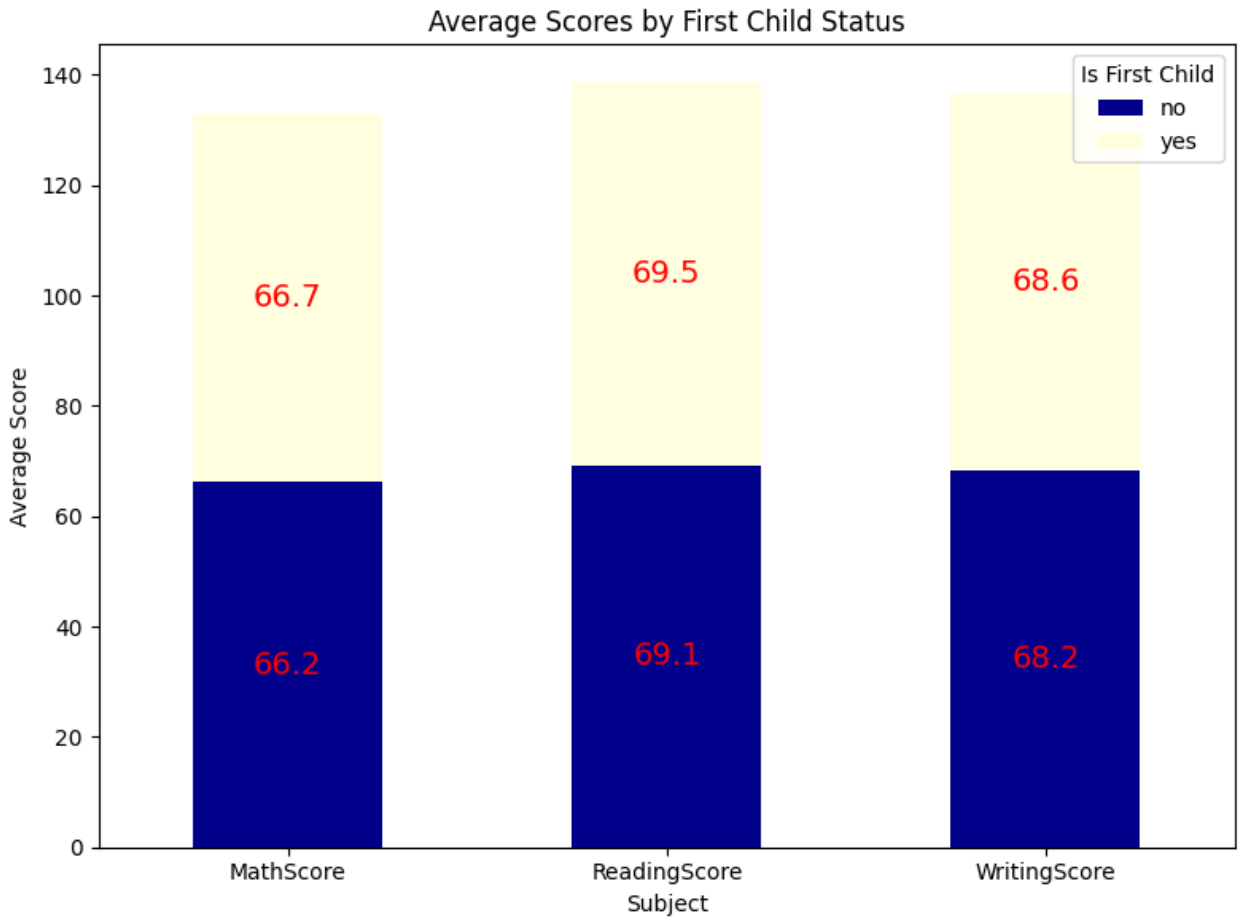
In [92]:
```
avg_scores_firstchild = df.groupby('IsFirstChild')[['MathScore', 'Reading
avg_scores_firstchild_T = avg_scores_firstchild.T

fc = avg_scores_firstchild_T.plot(kind='bar', stacked=True, figsize=(8, 6
```

```
plt.title('Average Scores by First Child Status')
plt.ylabel('Average Score')
plt.xlabel('Subject')
plt.xticks(rotation=0)
plt.legend(title='Is First Child')

for container in fc.containers:
    fc.bar_label(container, label_type='center', fontsize=14,color='red')

plt.tight_layout()
plt.show()
```



Insights

- First-born students score **slightly higher** across all subjects compared to their siblings.
- The differences are modest, with **reading scores** showing the largest gap (0.4 points).
- This may be due to **greater parental focus or academic expectations** for first-borns.

Recommendations

- Provide **equitable support** to all children in multi-child households.
- Encourage parents to **engage consistently** with each child's

education, regardless of birth order.

- Schools may consider **family counseling** programs to promote balanced academic nurturing.

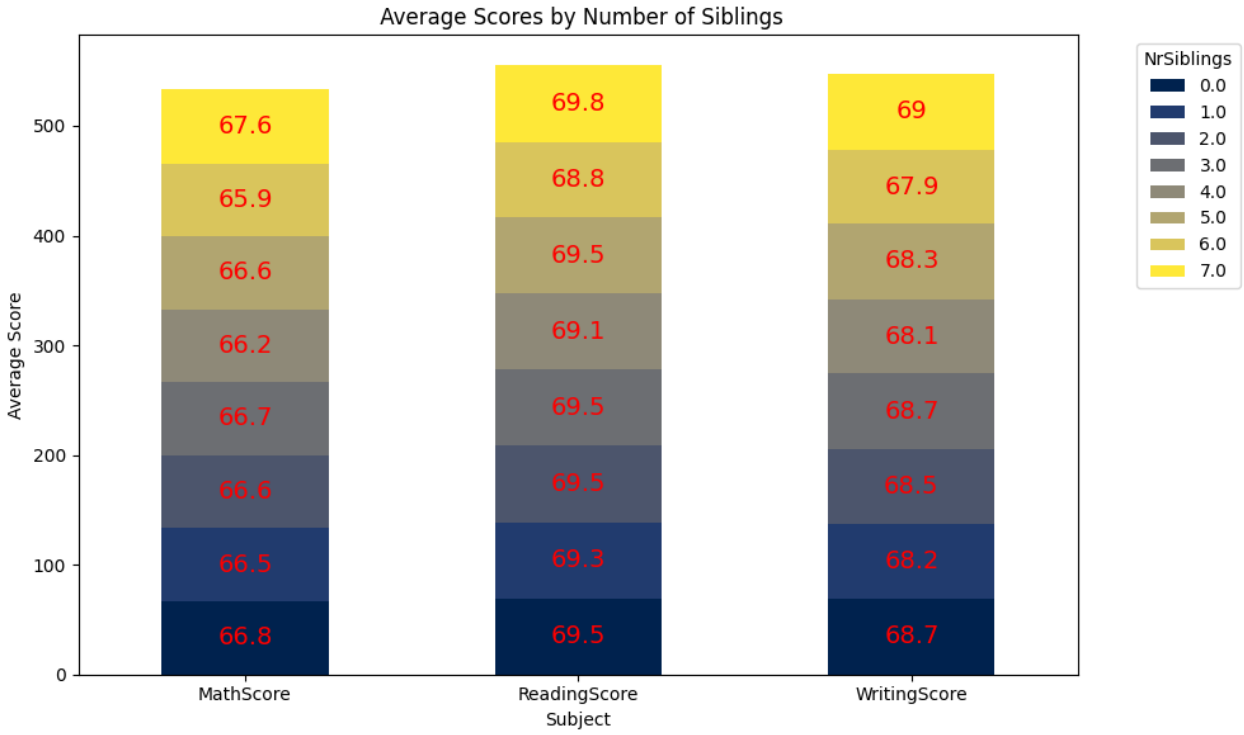
10 Impact Of Number Of Siblings On Scores

In [91]:
```
avg_scores_siblings = df.groupby('NrSiblings')[['MathScore', 'ReadingScor
avg_scores_siblings_T = avg_scores_siblings.T

ns = avg_scores_siblings_T.plot(kind='bar', stacked=True, figsize=(10, 6)
plt.title(' Average Scores by Number of Siblings')
plt.ylabel('Average Score')
plt.xlabel('Subject')
plt.xticks(rotation=0)
plt.legend(title='NrSiblings', bbox_to_anchor=(1.05, 1), loc='upper left'

for container in ns.containers:
    ns.bar_label(container, label_type='center', fontsize=14,color='red')

plt.tight_layout()
plt.show()
```



Insights

- Students with **0 or 1 sibling** tend to achieve **higher average scores** across all three subjects (Math, Reading, Writing).
- As the **number of siblings increases**, average scores show a **declining trend**, especially after 3 or more siblings.

- The **ReadingScore** remains relatively stable even as the number of siblings increases, showing more resilience compared to Math and Writing.
- Students with **6 or 7 siblings** have noticeably **lower performance** in Math and Writing compared to those with fewer siblings.

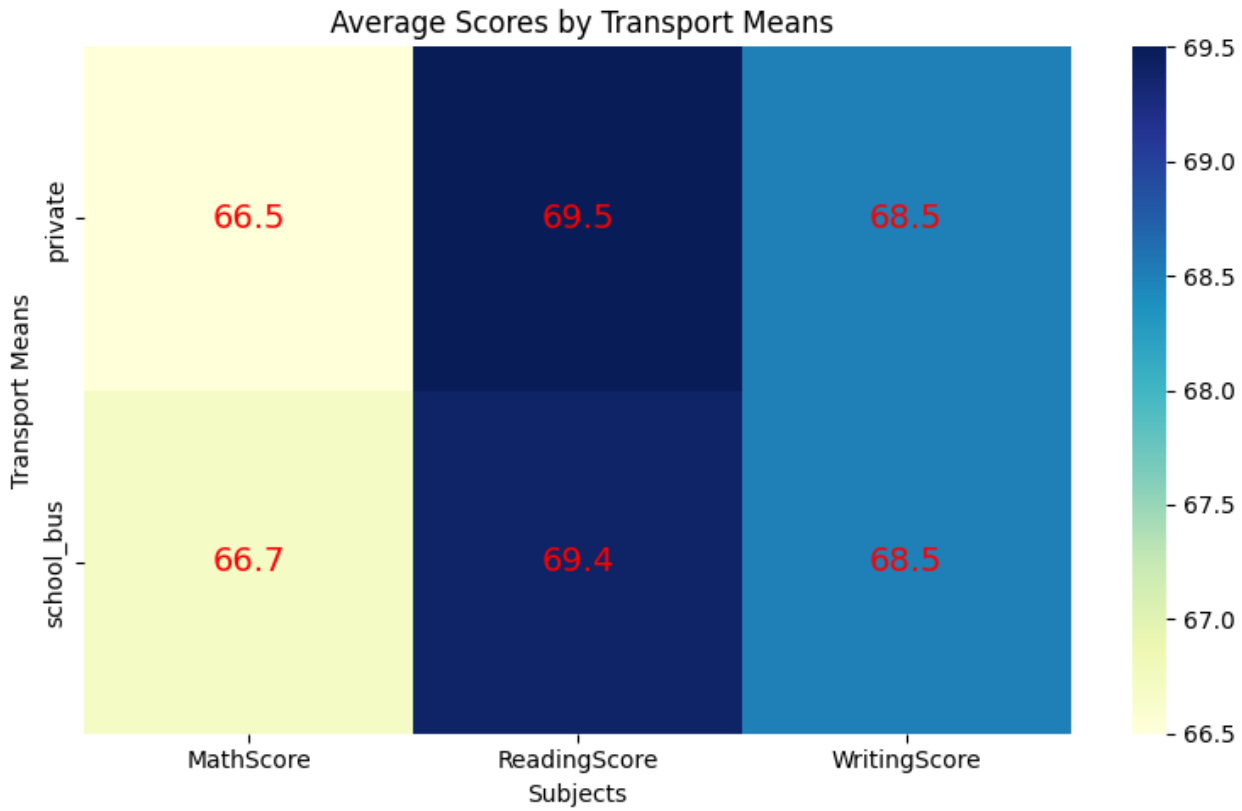
Recommendations

- Provide **targeted academic support** for students from larger families (4+ siblings), who may face limited individual attention or resources at home.
- Focus intervention efforts more on **Math and Writing**, where the performance drop is more pronounced.
- Consider **mentorship and peer tutoring programs** to help balance disparities in academic performance linked to sibling count.
- Explore **additional data** (e.g., socioeconomic status, parental involvement) to understand the broader context influencing these trends.

11.Imapct Of Transport Means On Academic Performance

In [94]:
```
heat_data = df.groupby('TransportMeans')[['MathScore', 'ReadingScore', 'W

plt.figure(figsize=(8, 5))
sns.heatmap(heat_data, annot=True, cmap="YlGnBu", fmt=".1f",annot_kws={"s
plt.title(" Average Scores by Transport Means")
plt.ylabel("Transport Means")
plt.xlabel("Subjects")
plt.tight_layout()
plt.show()
```



Insights

- Students using **school buses and private transport** show **almost identical performance** across all subjects.
- A **slightly higher reading score** is seen in those using private transport (by 0.1), but the differences are negligible.
- This suggests that **mode of transport alone may not be a strong factor** affecting academic performance.

Recommendations

- Focus on **ensuring punctuality and safety** of all transport modes rather than linking them to academic outcomes.
- Schools may monitor **time spent commuting** as a deeper factor if significant score differences emerge.

Final Summary: Factors Affecting Student Performance

This analysis explored how various demographic and lifestyle factors influence academic performance across Math, Reading, and Writing scores. Key insights and actionable recommendations are summarized below:

Key Insights

- **Gender**: Female students consistently scored higher in reading and writing; males had a slight edge in math.
- **Ethnic Group**: Group E had the highest average scores, highlighting potential cultural or support differences.
- **Parental Education**: Students whose parents had a master's or bachelor's degree performed better overall.
- **Lunch Type**: Standard lunch was strongly associated with higher scores compared to free/reduced lunch.
- **Test Preparation**: Students who completed test prep scored significantly higher in all subjects.
- **First Child & Siblings**: First-borns slightly outperformed others. Too many or zero siblings showed slightly lower scores.
- **Transport Means**: Little to no difference between private and school transport in terms of scores.
- **Study Hours**: Students studying more than 10 hours weekly consistently scored higher.
- **Sports Practice**: Regular physical activity showed a small positive correlation with scores.

Recommendations

- Encourage **test preparation programs** and ensure **equitable access**.
- Promote **healthy lunches**, especially for those receiving free/reduced meals.
- Support **family engagement** in academics regardless of parental education or child order.
- Encourage **balanced study schedules** and **physical activity** as both support better outcomes.
- Use these insights to guide **data-driven educational policies** and student support initiatives.