Data preprocessing and EDA

Mohamed Yessine Baananou

Ecole Polytechnique de Sousse

Utilizing Python Libraries for Data preprocessing and an EDA

& Data Vis

Professor: Dr. Ing. Rihab LAAJILI

2023 / 2024

[1. Introduction 4](#_Toc154852260)

[2. Data Overview 4](#_Toc154852261)

[3. Score Calculations and Grading 4](#_Toc154852262)

[4. Gender Analysis 5](#_Toc154852263)

[5. Race/Ethnicity Analysis 5](#_Toc154852264)

[6. Parental Level of Education Analysis 5](#_Toc154852265)

[7. Lunch Analysis 6](#_Toc154852266)

[8. Test Preparation Course Analysis 6](#_Toc154852267)

[9. Score Distribution 6](#_Toc154852268)

[10. Grade Distribution 7](#_Toc154852269)

[11. Exam Status Analysis 7](#_Toc154852270)

[12. Gender vs. Score & Grade 8](#_Toc154852271)

[13. Race/Ethnicity vs. Score & Grade 8](#_Toc154852272)

[14. Lunch vs. Score & Grade 9](#_Toc154852273)

[15. Test Preparation vs. Score & Grade 9](#_Toc154852274)

[16. Parent Education vs. Score & Grade 10](#_Toc154852275)

[17. Additional Visualizations 10](#_Toc154852276)

[18. Race/Ethnicity vs. Parent Education & Lunch 12](#_Toc154852277)

[19. Conclusion 13](#_Toc154852278)

# Introduction

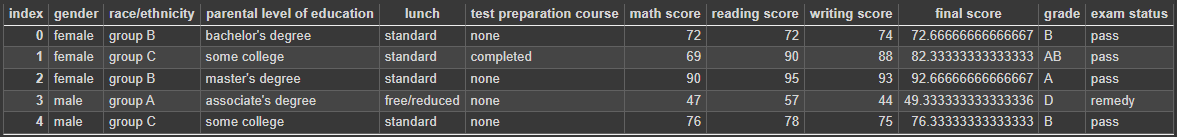
The analysis is performed on the ['Students Performance in Exams'](https://www.kaggle.com/datasets/spscientist/students-performance-in-exams) dataset to gain insights into the academic performance of students. The dataset includes information about students' gender, race/ethnicity, and parental level of education, lunch type, test preparation course, and scores in math, reading, and writing.

# Data Overview

* The dataset is loaded into a Pandas Data Frame, and the first few rows are displayed to provide an initial overview.
* Missing values in each column are checked and found to be absent.
* The unique values and counts for categorical columns (gender, race/ethnicity, parental level of education, lunch, and test preparation course) are explored.

# Score Calculations and Grading

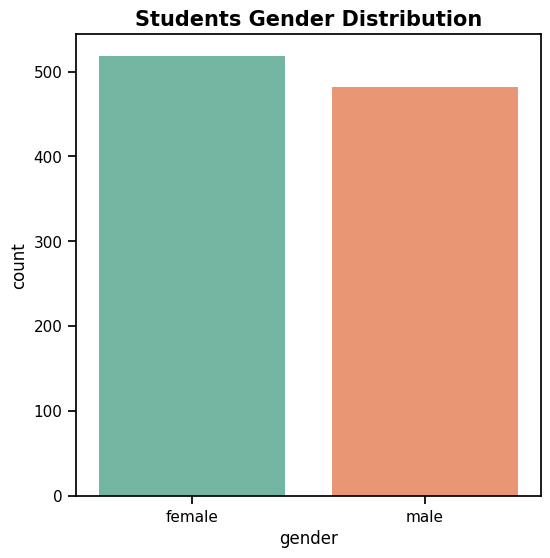
* A 'final score' is calculated as the average of math, reading, and writing scores for each student.
* Students are assigned grades ('A' to 'F') based on their final scores.
* An 'exam status' is assigned to each student, indicating whether they passed, need remediation, or failed.



# Gender Analysis

* Gender distribution is visualized using a count plot.
* Total male and female students are displayed.

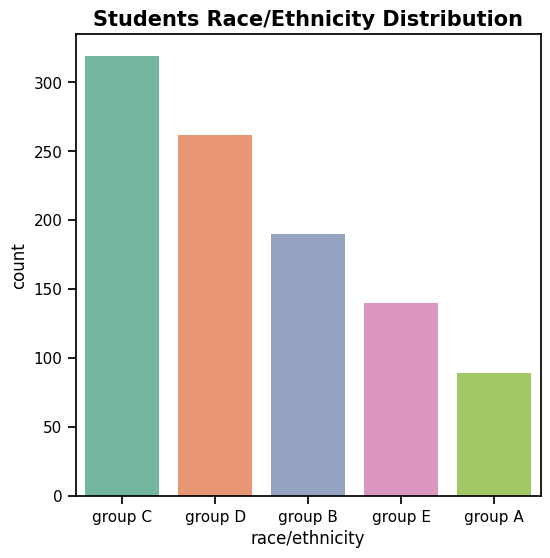
**Analysis: The gender distribution in the dataset is relatively balanced, with a slightly higher number of female students.**



# Race/Ethnicity Analysis

Race/ethnicity distribution is visualized, and the count of students in each category is displayed.

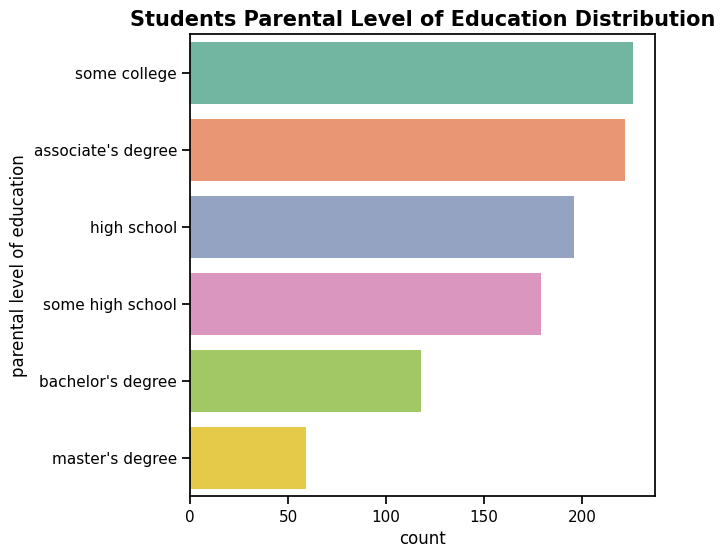
**Analysis: The dataset has a diverse representation of students across different race/ethnicity groups, with Group C being the most populous.**



# Parental Level of Education Analysis

Distribution of parental level of education is visualized, and the count of students in each category is displayed.

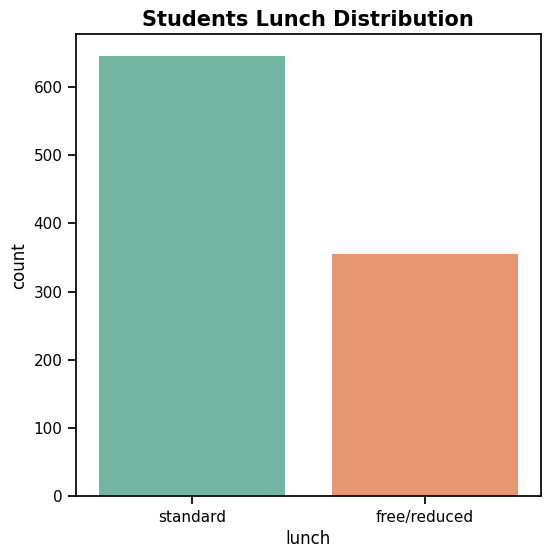
**Analysis: The parental level of education varies, with a substantial number of students having education beyond high school.**



# Lunch Analysis

Lunch distribution is visualized, and the count of students with standard and free/reduced lunch is displayed.

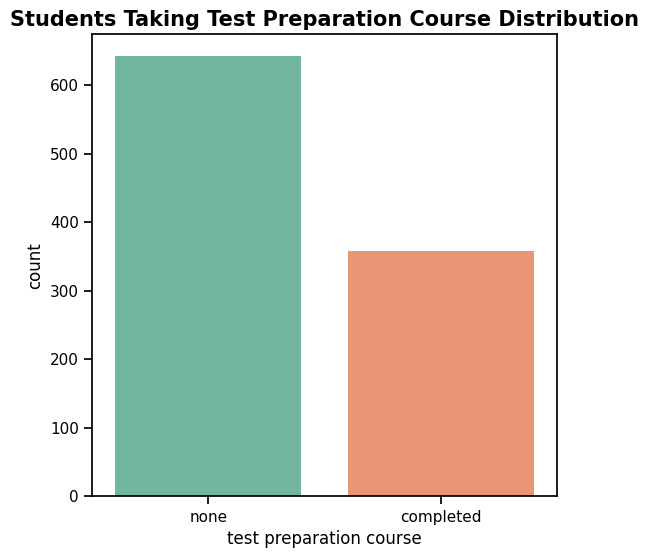
**Analysis: A majority of students have a standard lunch, indicating potential socioeconomic differences that may impact performance.**



# Test Preparation Course Analysis

Test preparation course distribution is visualized, and the count of students taking and not taking the course is displayed.

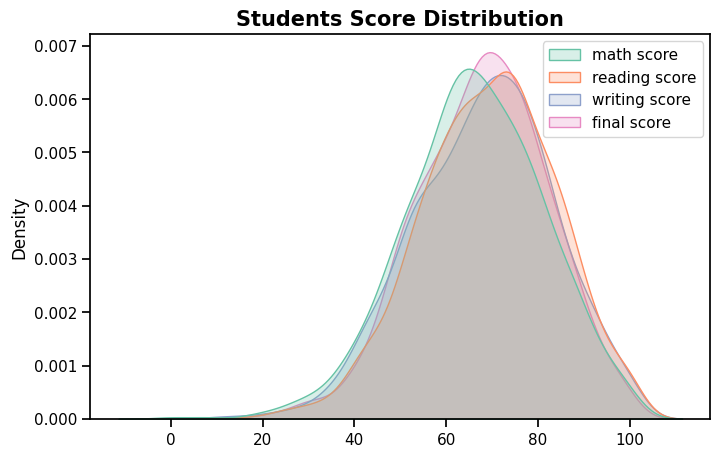
**Analysis: A significant portion of students did not take the test preparation course, suggesting potential differences in preparedness for exams.**



# Student Score Distribution

The overall score distribution is visualized using a kernel density plot, Average scores for math, reading, writing, and final score are displayed.

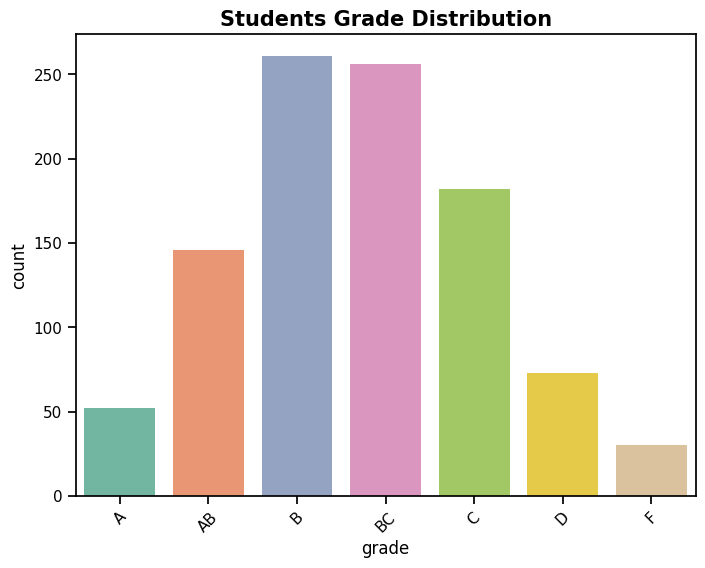
**Analysis: The average scores across all subjects and the final score are relatively balanced.**



# Students Grade Distribution

If the 'grade' column exists, the grade distribution is visualized, and the count of students with each grade is displayed.

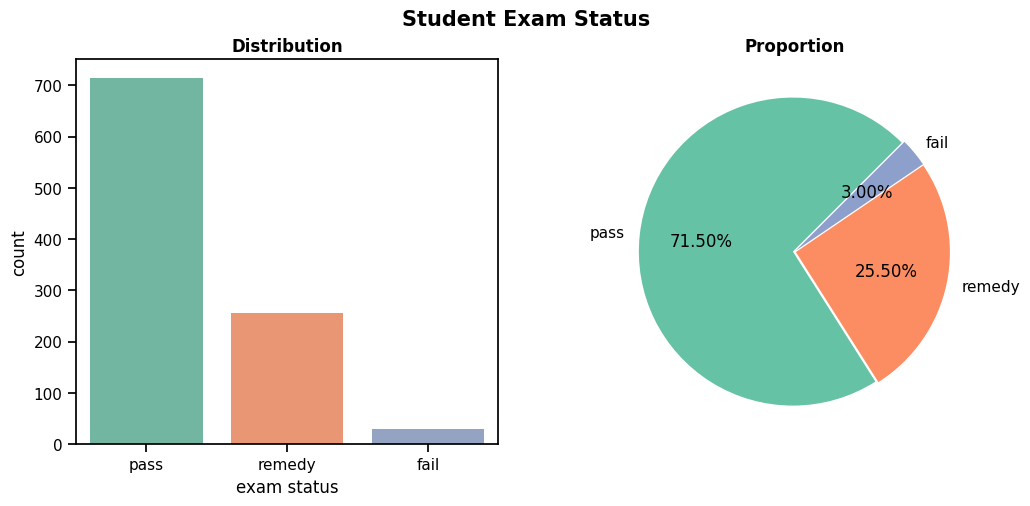
**Analysis: The majority of students fall within the 'B' and 'C' grade categories.**



# Student Exam Status

* Exam status distribution is visualized using a count plot.
* Proportional distribution of exam status is visualized in a pie chart.
* Total students in each exam status category are displayed.

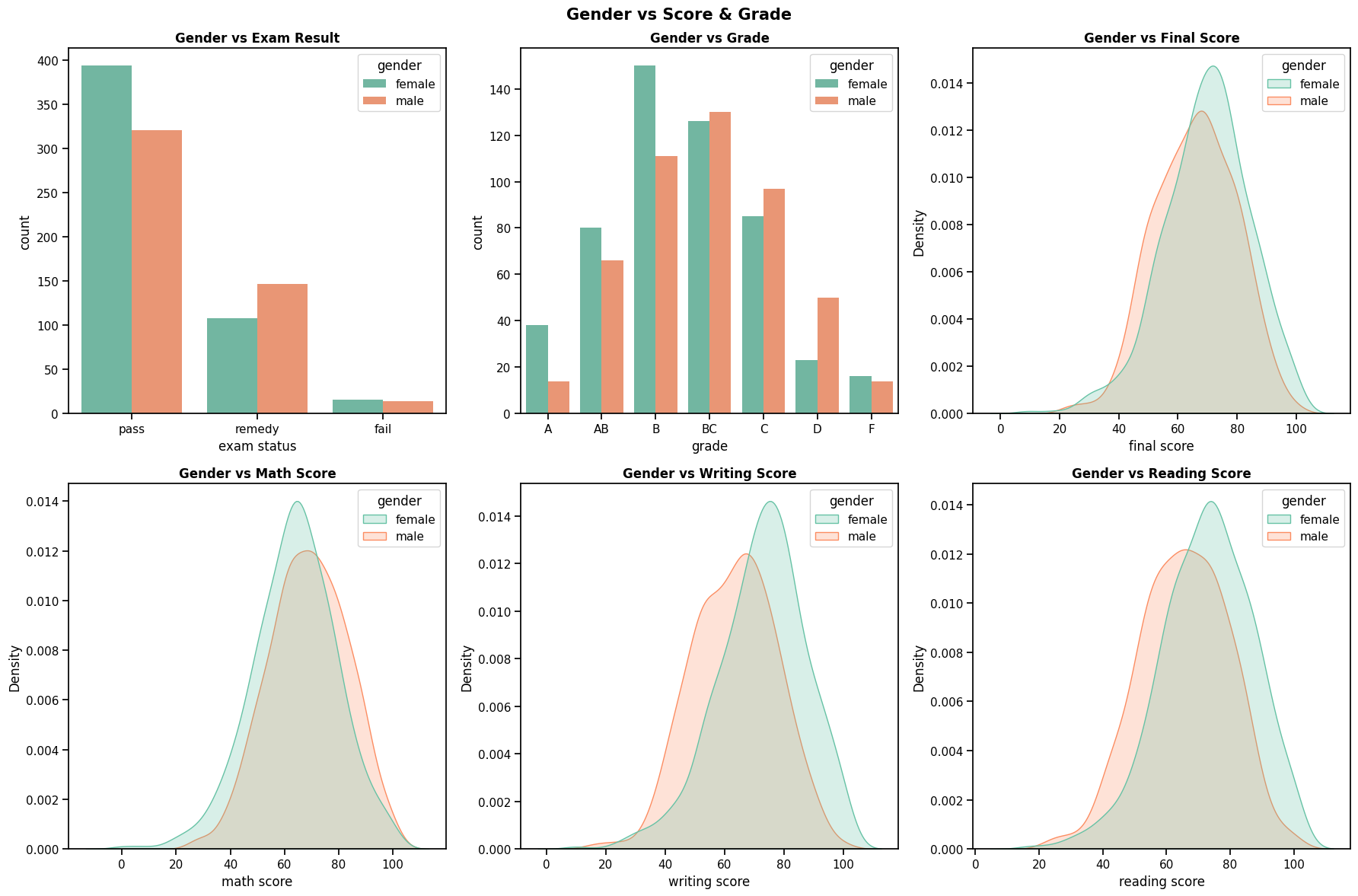
**Analysis: A significant majority of students have passed the exams, while a smaller portion requires remediation, and a very few have failed.**



# Gender vs. Score & Grade

Relationships between gender and exam result, grade, and scores are visualized.

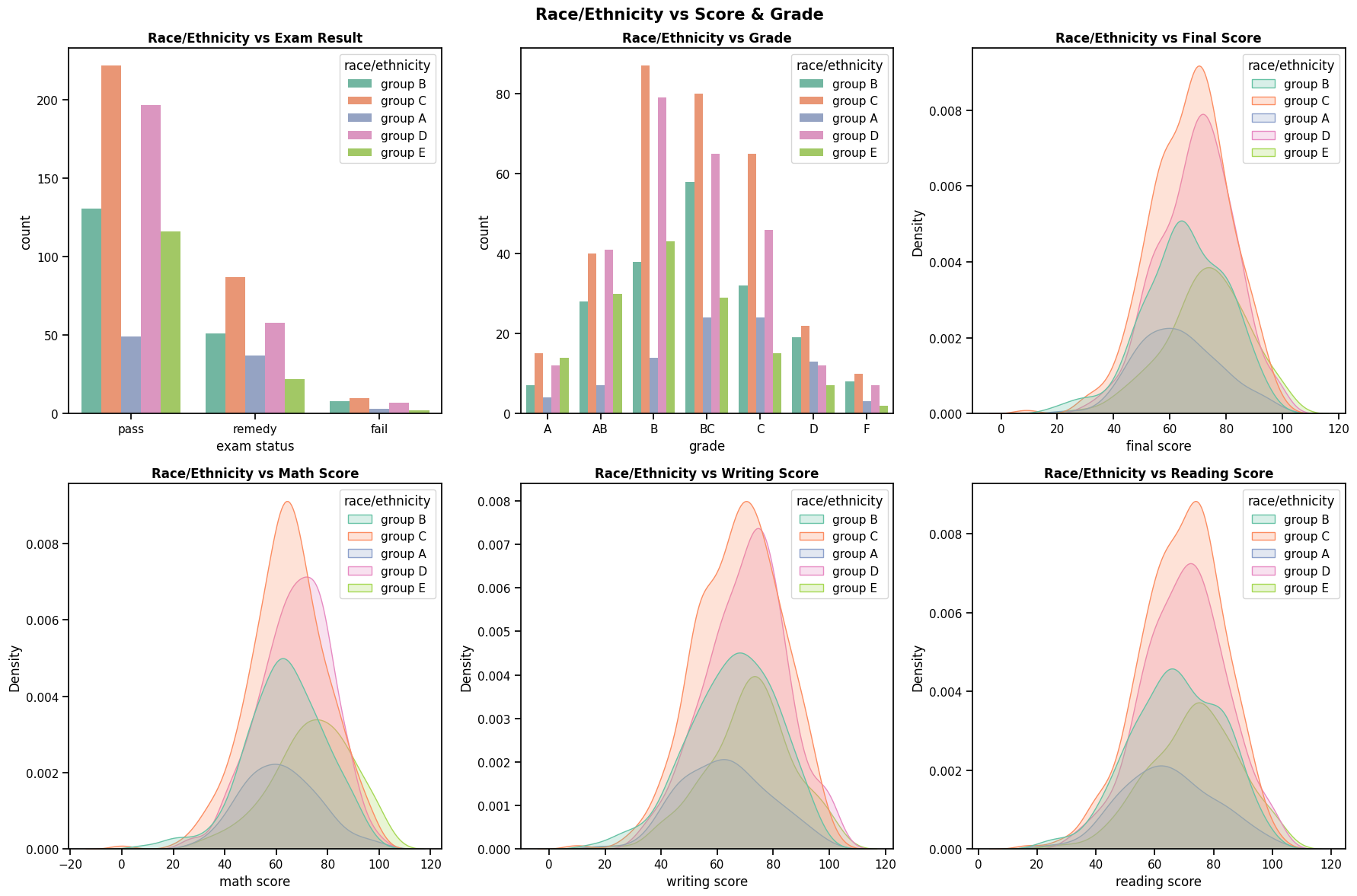
**Analysis: This section visually explores the relationship between gender and exam performance, including scores and grades.**



# Race/Ethnicity vs. Score & Grade

Relationships between race/ethnicity and exam result, grade, and scores are visualized.

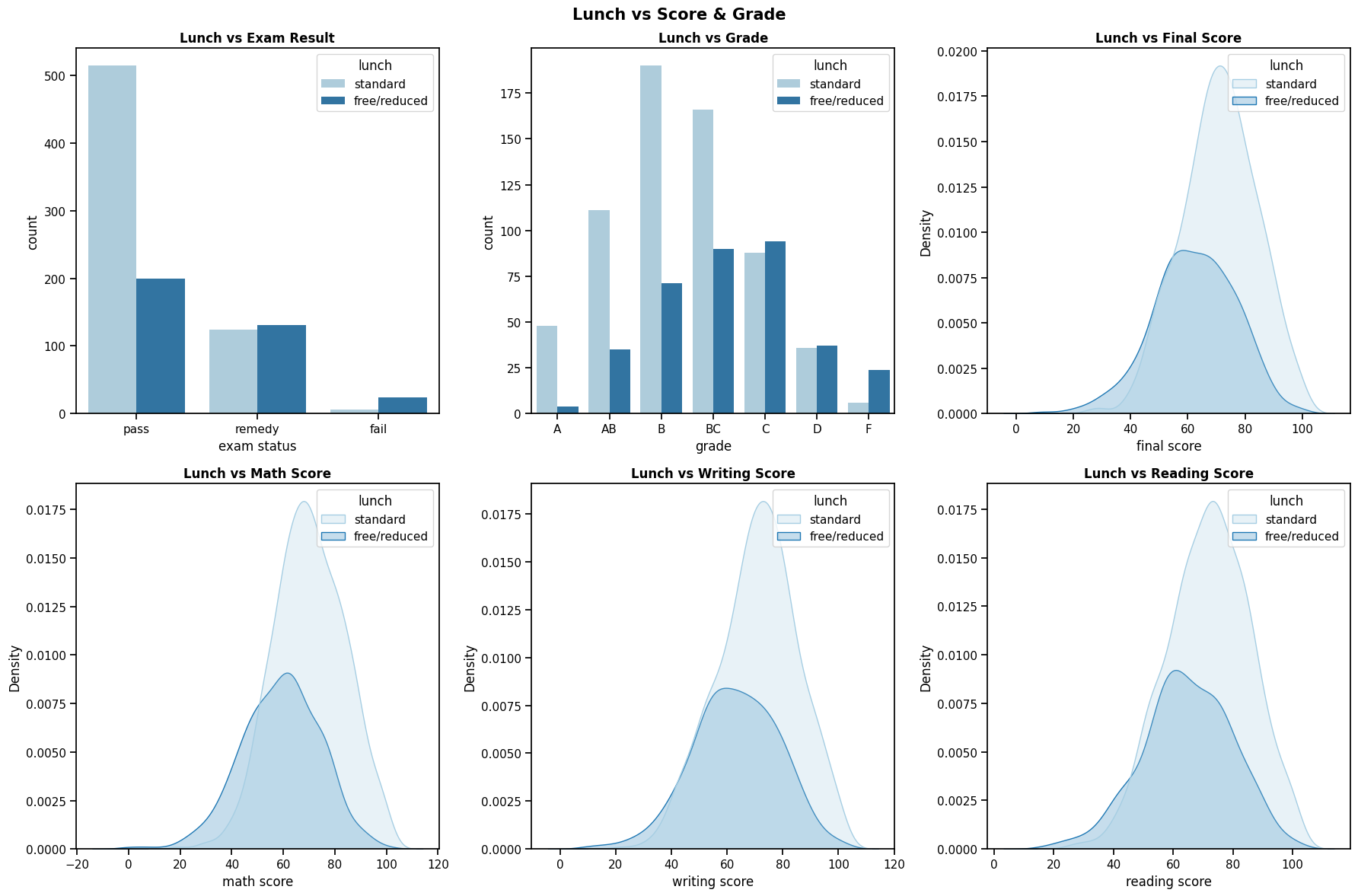
**Analysis: Examines how race/ethnicity correlates with scores and grades, considering potential variations.**



# Lunch vs. Score & Grade

Relationships between lunch type and exam result, grade, and scores are visualized.

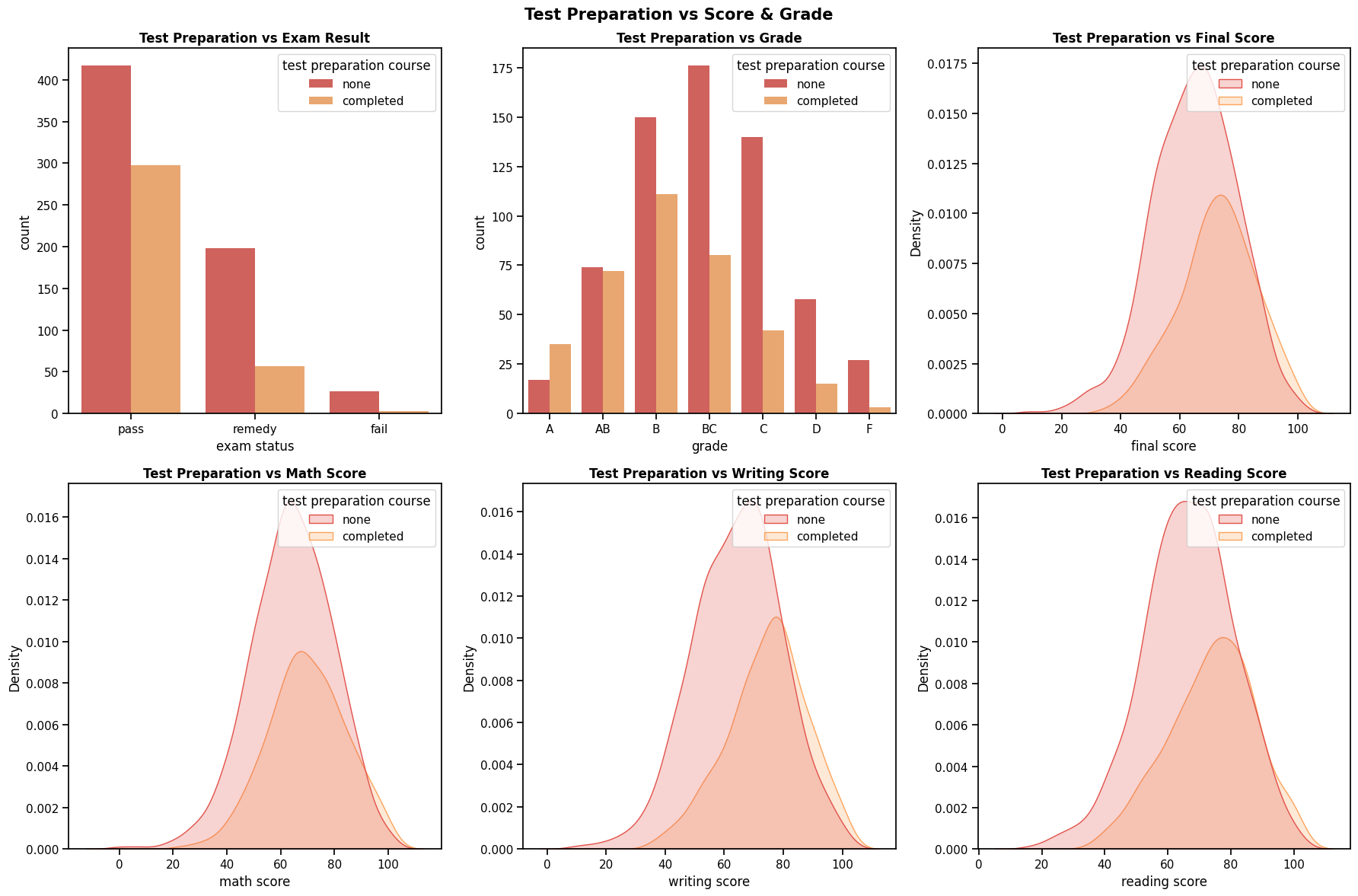
**Analysis: Investigates whether the type of lunch has an impact on exam scores and grades.**



# Test Preparation vs. Score & Grade

Relationships between test preparation and exam result, grade, and scores are visualized.

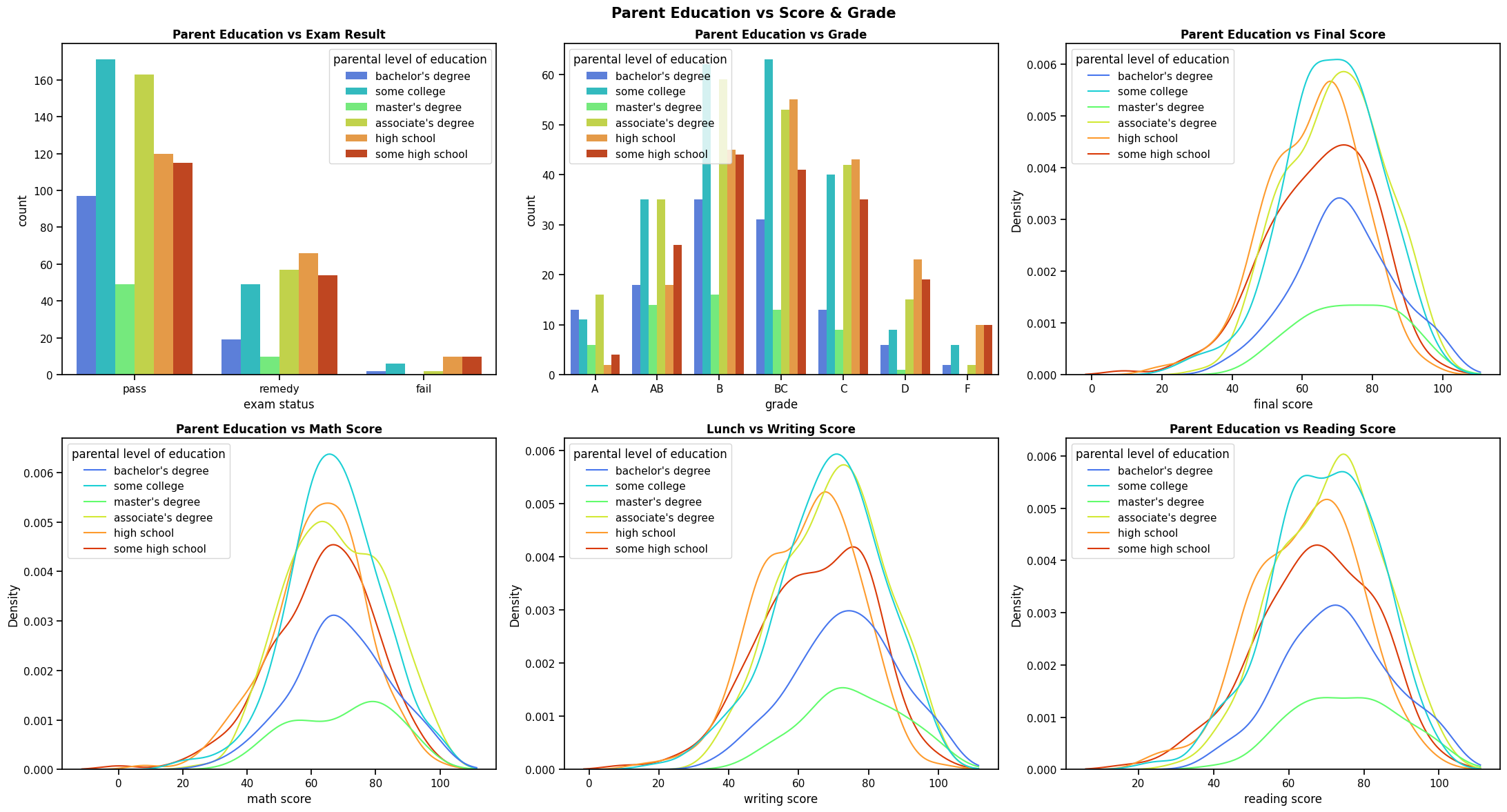
**Analysis: Explores how completion of the test preparation course influences exam scores and grades.**



# Parent Education vs. Score & Grade

A visualization of relationships between parental education and exam result, grade, and scores

**Analysis: Analyzes the relationship between parental education and exam performance, considering scores and grades.**

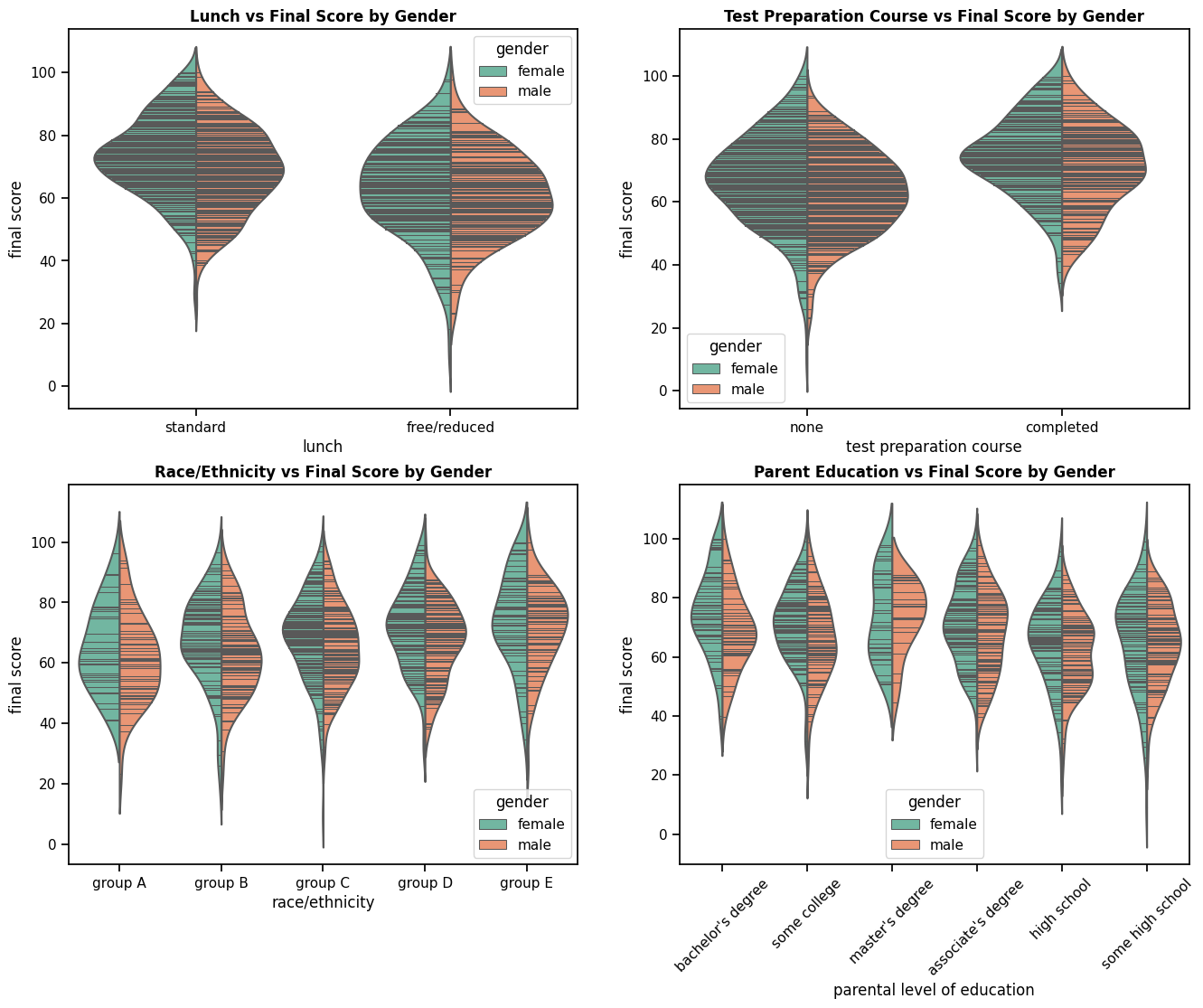


# Additional Visualizations

* 1. Lunch vs. Final Score (by Gender)

Investigating whether lunch type influences final scores, divided by gender.

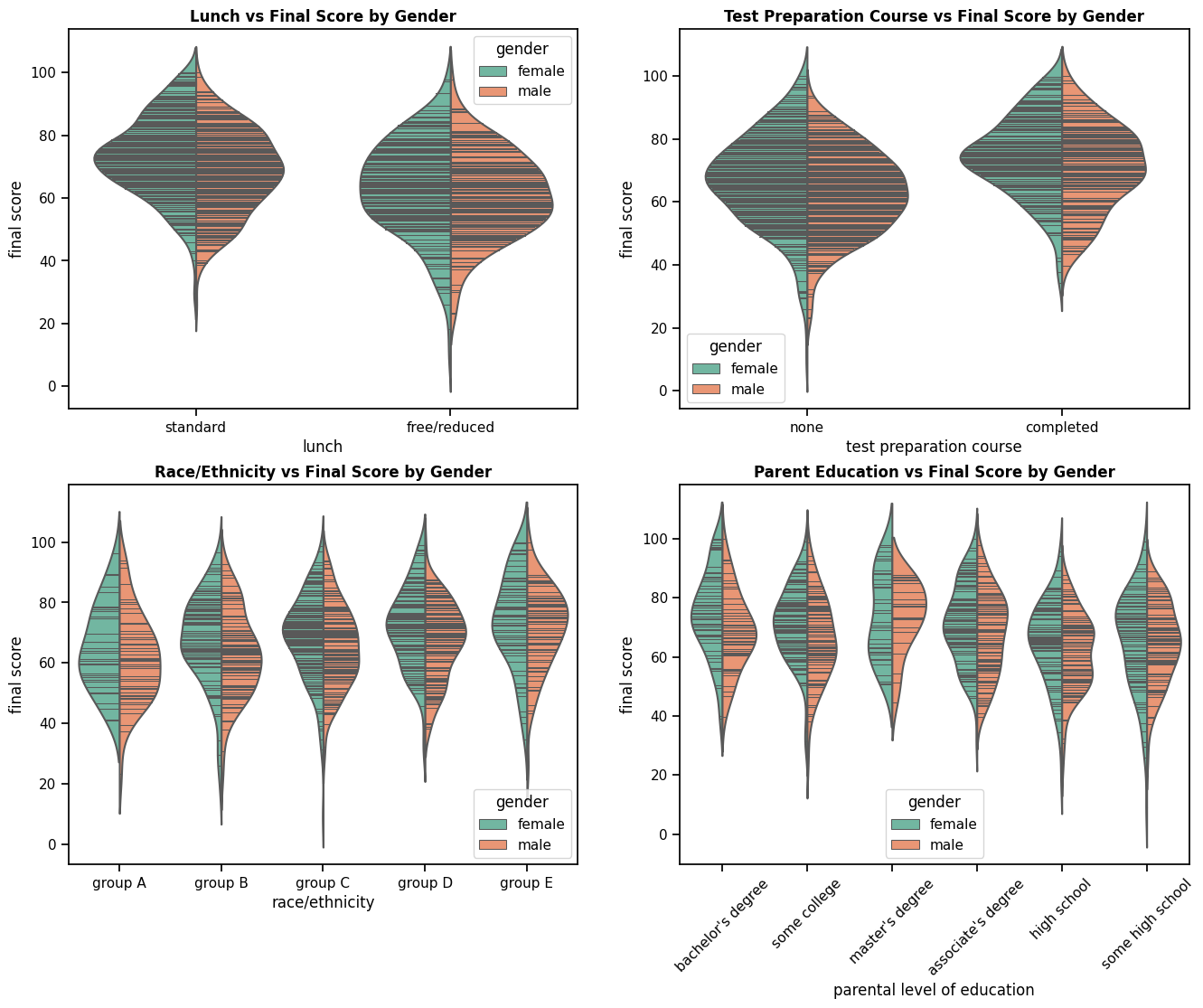
**Analysis: Explores how lunch type and gender together influence final scores using violin plots.**



* 1. Test Preparation vs. Final Score (by Gender)

Exploring the impact of test preparation on final scores, with a focus on gender differences.

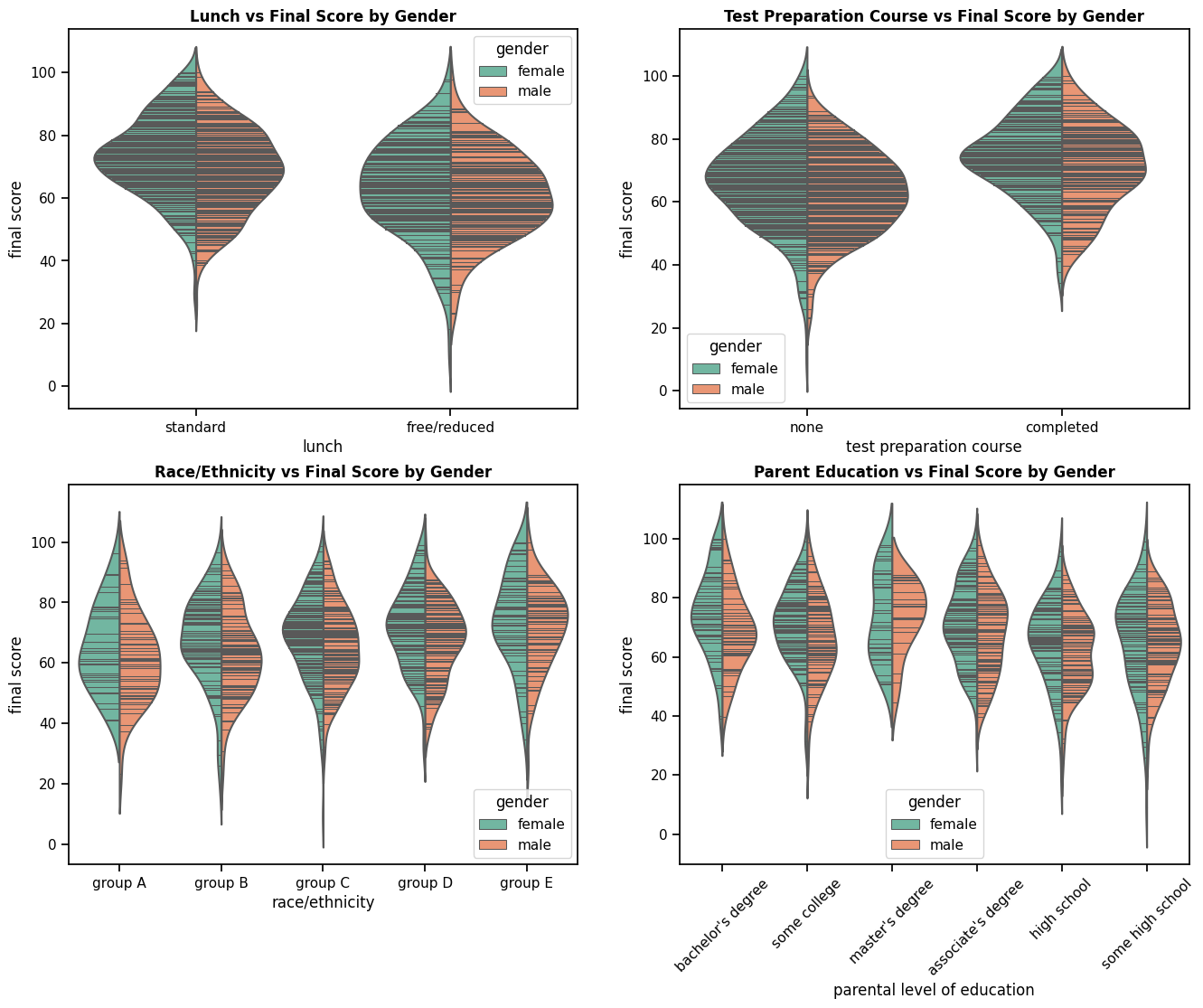
**Analysis: Investigates the impact of test preparation on final scores, considering gender differences.**



* 1. Race/Ethnicity vs. Final Score (by Gender)

Examining how race/ethnicity correlates with final scores, considering gender distinctions.

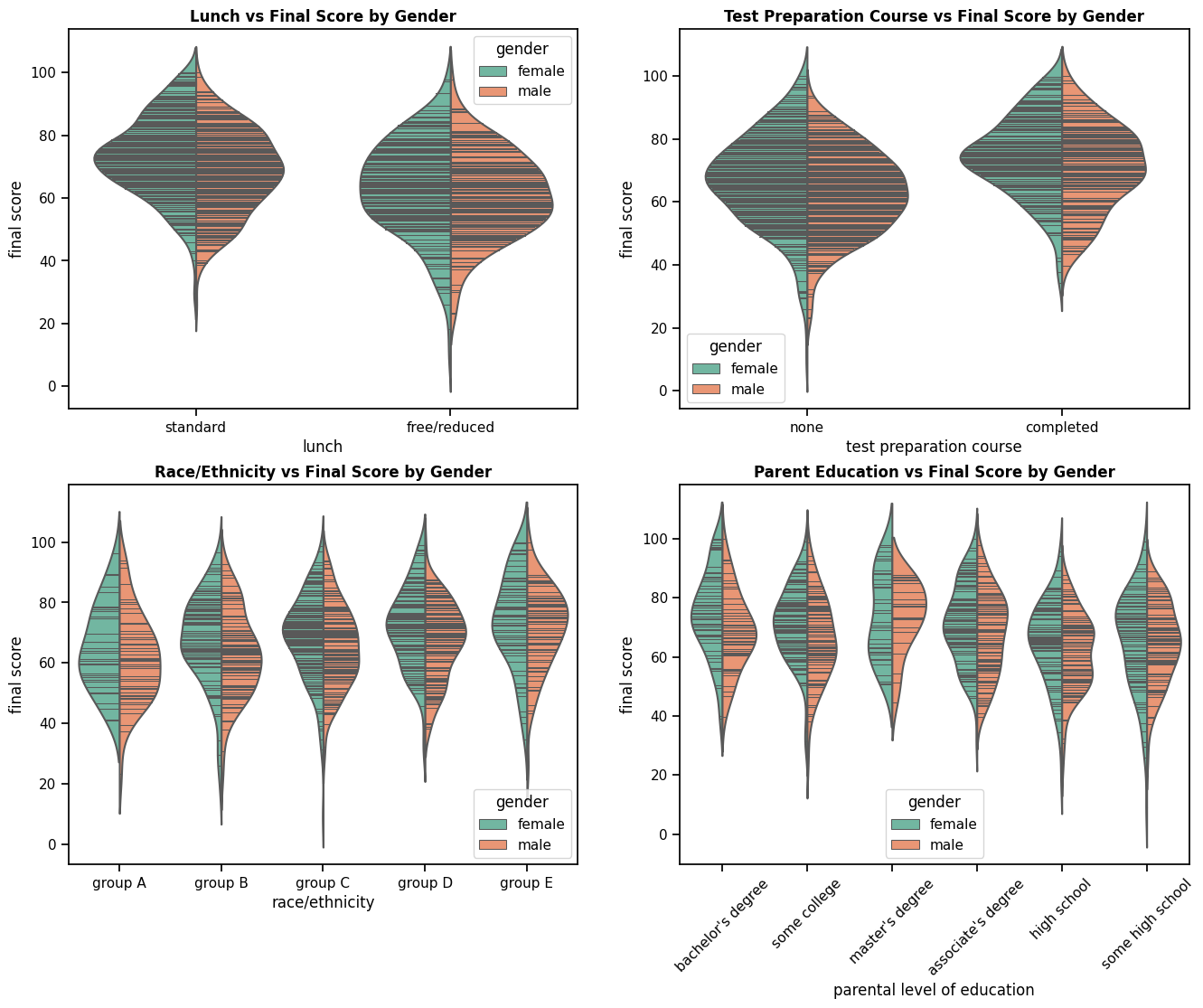
**Analysis: Examines how race/ethnicity, in conjunction with gender, correlates with final scores.**



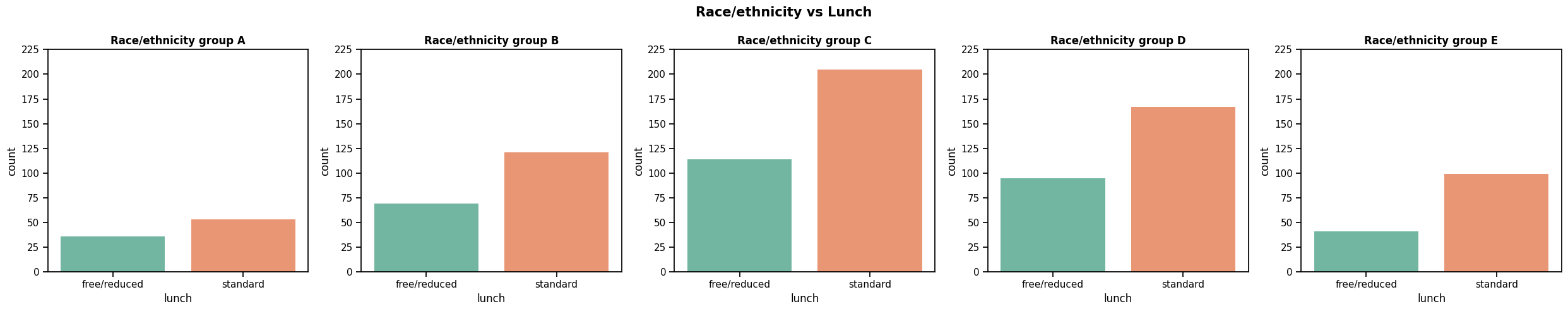
* 1. Parental Education vs. Final Score (by Gender)

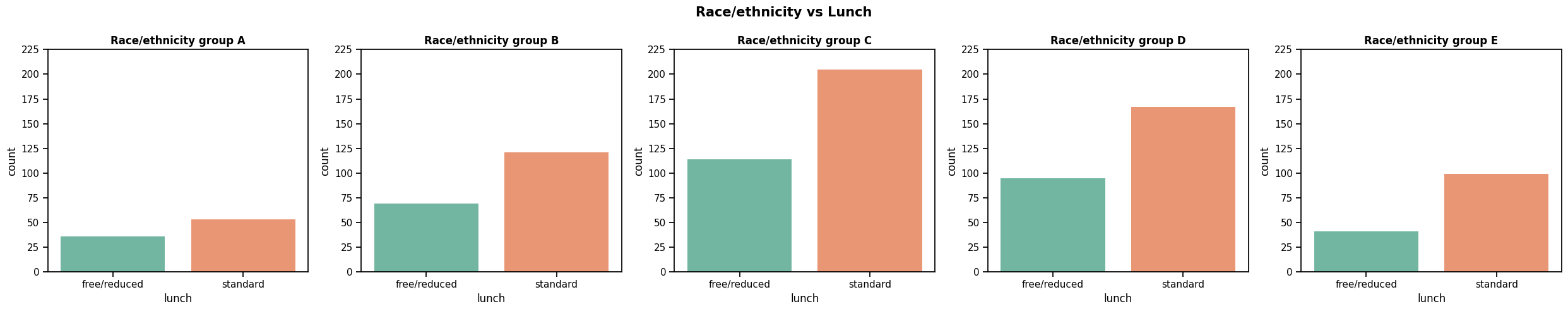
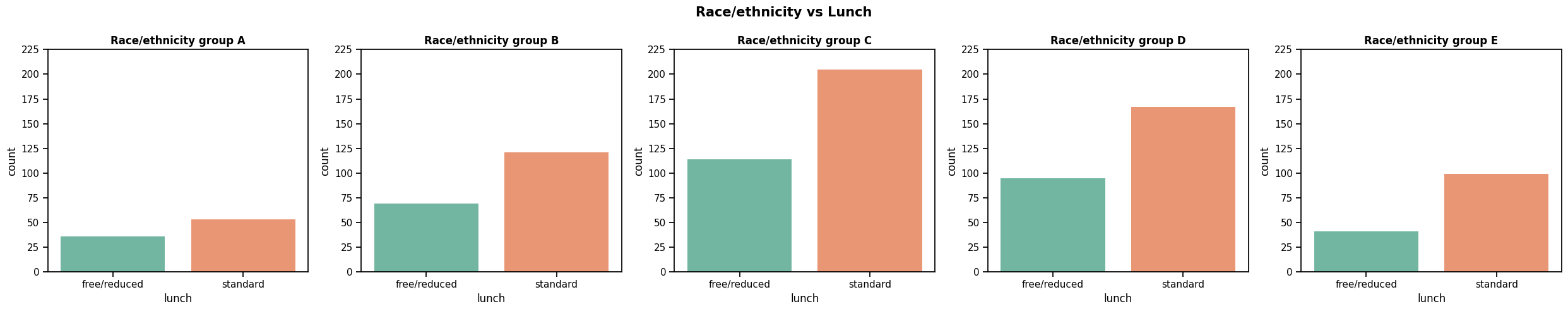
Understanding the relationship between parental education and final scores, while observing gender nuances.

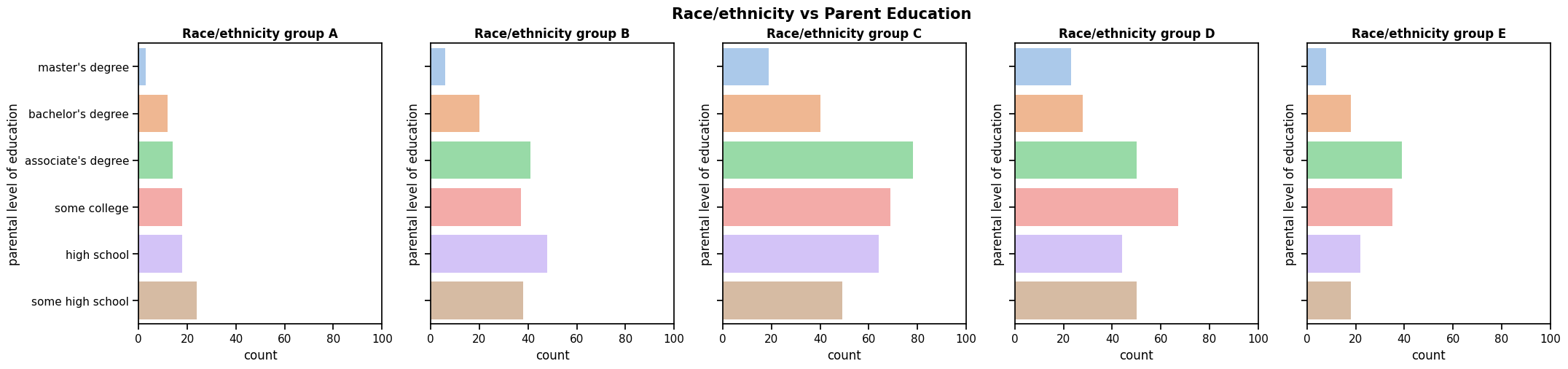
**Analysis: Analyzes the relationship between parental education, gender, and final scores using violin plots.**

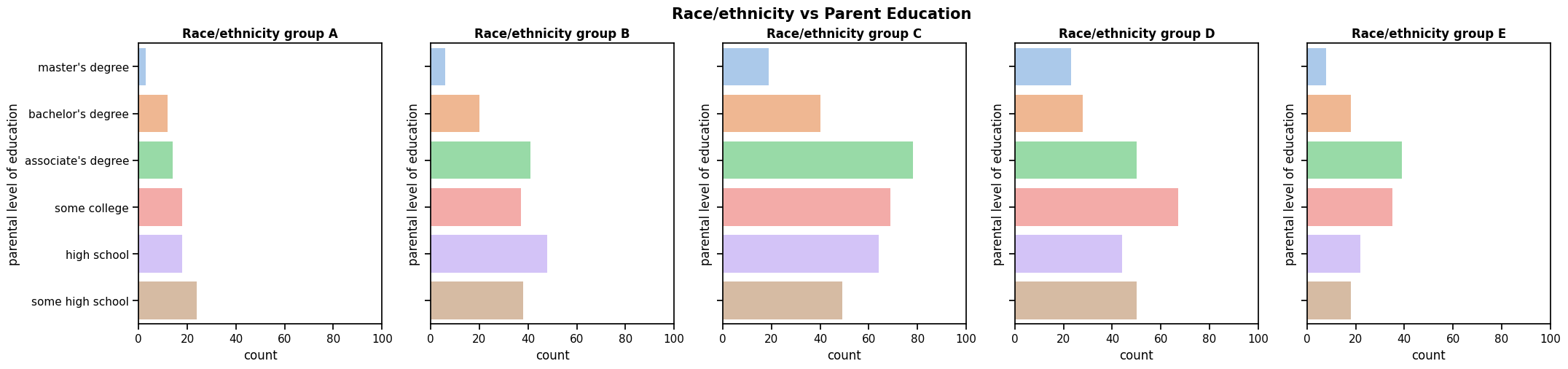


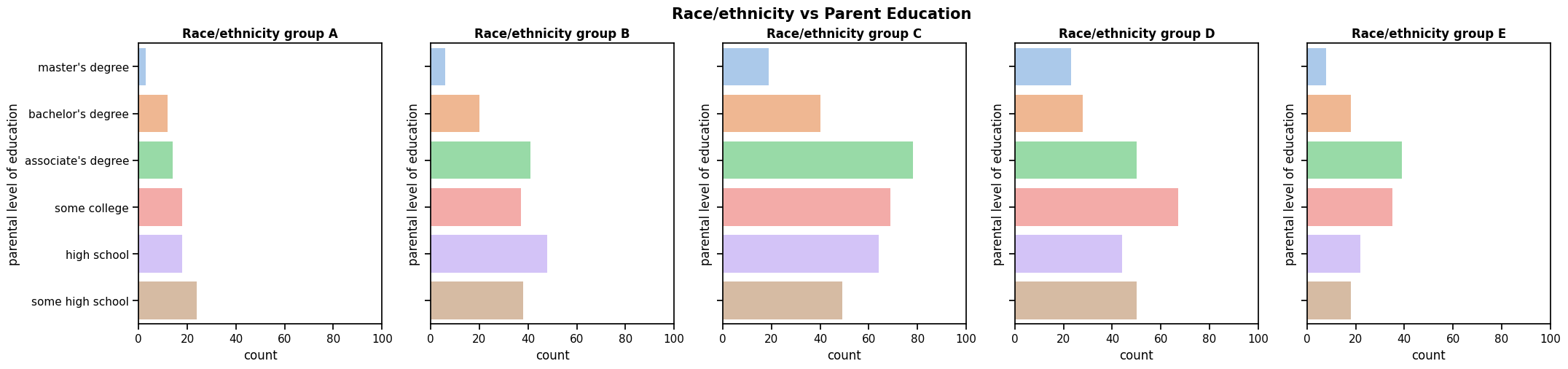
# Race/Ethnicity vs. Parent Education & Lunch

Count plots depict the relationship between race/ethnicity and parental level of education, as well as lunch type.









# Conclusion

* The dataset includes a diverse group of students with balanced gender representation.
* Parental education levels vary, with a considerable number of students having education beyond high school.
* Most students have a standard lunch, suggesting potential socioeconomic disparities.
* A significant portion of students did not take the test preparation course, indicating varied levels of preparedness.
* The overall average scores across subjects are balanced.
* The majority of students fall within the 'B' and 'C' grade categories.
* A substantial number of students have passed the exams, while a smaller portion requires remediation, and very few have failed.
* Additional visualizations explore the relationships between gender, race/ethnicity, parental education, lunch type, and test preparation with exam scores and grades.
* In conclusion, the analysis provides a comprehensive understanding of student performance, highlighting potential factors influencing academic outcomes. Further, it sheds light on disparities and patterns within different demographic groups, aiding in informed decision-making for educational interventions and support.