

---

# CAUSAL DISCOVERY REPORT ON CLEANED STUDENTS PERFORMANCE

---

## TECHNICAL REPORT



April 8, 2025

## ABSTRACT

This report presents a causal discovery analysis of a dataset containing various attributes related to student performance, including demographics, socioeconomic factors, and academic metrics in math, reading, and writing. Using the PC algorithm, we explored the causal relationships among these variables, revealing significant influences of factors like gender, race/ethnicity, and lunch type on academic performance. Our findings indicate that gender directly impacts all assessed scores while race/ethnicity affects math scores, suggesting disparities in educational outcomes tied to demographic characteristics. Furthermore, lunch type was identified as a crucial socioeconomic factor influencing academic results, with test preparation courses enhancing performance in reading and writing. The integration of causal discovery with a bootstrap method provided a robust framework for evaluating the reliability of causal edges, while our analysis highlighted the need for targeted educational interventions. This study contributes valuable insights for educators and policymakers aiming to understand and address the multifaceted influences on student achievement.

**Keywords** Causal Discovery, Large Language Model, PC, Cleaned Students Performance

## 1 Introduction

In recent years, the field of causal discovery and inference has gained prominence, particularly in educational settings where understanding the factors influencing student performance is crucial. This report focuses on a dataset that encompasses various attributes pertaining to students, including demographic information such as gender and race/ethnicity, socioeconomic indicators like lunch type, and academic performance metrics across math, reading, and writing. The relationships hypothesized among these variables underscore the multifaceted nature of educational outcomes, suggesting that parental education levels, socioeconomic status, cultural influences, and preparation courses play instrumental roles in shaping academic success. By applying causal discovery techniques to this dataset, we aim to uncover the underlying causal structures and relationships that elucidate how these variables interact to impact student performance, thereby providing valuable insights for educators and policymakers alike.

## 2 Background Knowledge

### 2.1 Detailed Explanation about the Variables

The dataset comprises various variables that capture student demographics and academic performance. The **gender** variable identifies the sex of the student, which can impact academic achievement due to sociocultural influences. **Race/Ethnicity** classifies students based on their backgrounds, affecting access to educational resources. The **parental level of education** reflects the highest educational attainment of parents, known to influence student motivation and outcomes. The **lunch** variable serves as an indicator of socioeconomic status, with implications for nutrition and educational support. The dataset also includes test preparation and subject scores, such as **math score**, **reading score**, **writing score**, **total score**, and **average score**, which collectively offer measures of academic performance.

In terms of background knowledge, it is crucial to acknowledge that students from lower socioeconomic backgrounds may face challenges that impact their academic results. The effect of educational policies aimed at bridging achievement gaps is also significant. Additionally, cultural influences and psychological factors like motivation can shape student performance. Peer relationships can further affect student success, especially in collaborative settings. Incorporating these dimensions will enhance the understanding of the causal relationships in the dataset, making it easier to design effective causal discovery algorithms.

## 2.2 Possible Causal Relations found by LLM

The following are potential causal relationships suggested by the language model, which are visualized in Figure 1. Please note that only variables present in our dataset are included in the figure.

- **parental level of education → lunch:** Parents with higher education levels may be more likely to secure stable employment, leading to better economic conditions.

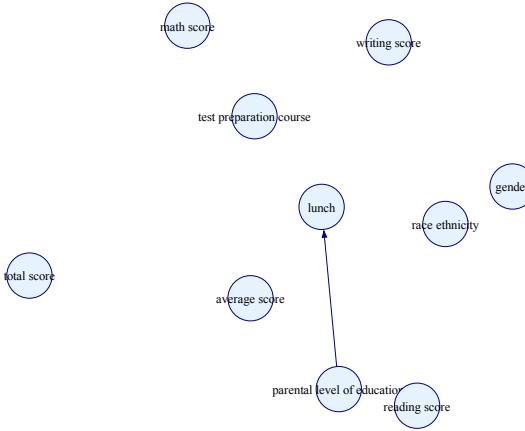


Figure 1: A Causal Graph Suggested by LLM.

## 3 Dataset Descriptions and EDA

The following provides a preview of our original dataset. If the dataset contains more than 10 columns, a random subset of 10 columns is displayed for illustrative purposes.

Table 1: Dataset Preview.

gender	race ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score	total score	average score
0	1	1	1	0	72	72	74	218	72.666667
0	2	4	1	1	69	90	88	247	82.333333
0	1	3	1	0	90	95	93	278	92.666667
1	0	0	0	0	47	57	44	148	49.333333
1	2	4	1	0	76	78	75	229	76.333333

### 3.1 Data Properties

We employed several statistical methods to identify data properties, including:

#### Basic Data Characteristics

The shape of the data, variable types, and the presence of missing values were assessed directly from the DataFrame. In contrast, properties such as time-series structure and heterogeneity were inferred with LLM based on user queries and DataFrame.

## Linearity Testing

We conducted the Ramsey's RESET test to assess linearity between each pair of variables. When the total number of possible variable pairs was fewer than 100, all pairs were tested. If the number exceeded 100, a random subset of 100 pairs was selected for testing to ensure computational feasibility. To account for multiple testing, we employed the Benjamini and Yekutieli procedure, which is robust when dealing with dependent or correlated data. The linearity assumption was considered satisfied only if all tested pairs exhibited linearity; otherwise, it was considered violated.

## Normality of Residuals

The assumption of Gaussian (normally distributed) noise was assessed using the Shapiro-Wilk test. The testing approach depended on the outcome of the linearity evaluation. If linearity was satisfied, we fitted ordinary least squares (OLS) models for each variable pair and extracted the residuals for testing. If linearity was not satisfied, we used a flexible non-parametric method—locally weighted scatterplot smoothing (LOWESS)—to model the relationships and obtain residuals. The Benjamini and Yekutieli correction was again applied to control for false discovery under multiple testing.

Properties of the dataset we analyzed are listed below.

Table 2: Data Properties.

Shape ( $n \times d$ )	Data Type	Missing Value	Linearity	Gaussian Errors	Time-Series	Heterogeneity
(1000, 10)	Mixture	False	False	False	False	False

## 3.2 Distribution Analysis

The following figure presents distributions of various variables. The orange dashed line indicates the mean, while the black solid line denotes the median. Variables are categorized into three types based on their distributional characteristics.

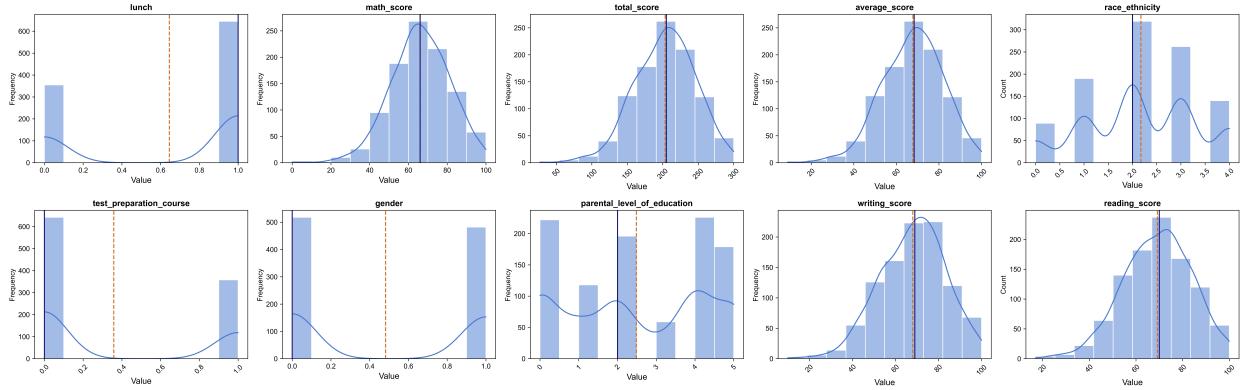


Figure 2: Distribution Plots of Variables.

## Numerical Variables

- Slight left skewed distributed variables: lunch, total score, average score, writing score, reading score
- Slight right skewed distributed variables: math score, race ethnicity, test preparation course, gender, parental level of education
- Symmetric distributed variables: None

### 3.3 Correlation Analysis

- **Strongly Correlated Variables ( $\geq 0.9$ ):** total score - math score, average score - math score, average score - total score, writing score - math score, writing score - total score, writing score - average score, reading score - math score, reading score - total score, reading score - average score, reading score - writing score
- **Moderately Correlated Variables (0.1 – 0.9):** None
- **Weakly Correlated Variables ( $\leq 0.1$ ):** math score - lunch, total score - lunch, average score - lunch, race ethnicity - math score, race ethnicity - total score, race ethnicity - average score, test preparation course - math score, test preparation course - total score, test preparation course - average score, gender - math score, gender - total score, gender - average score, writing score - lunch, writing score - race ethnicity, writing score - test preparation course, writing score - gender, reading score - lunch, reading score - race ethnicity, reading score - test preparation course, reading score - gender, etc.

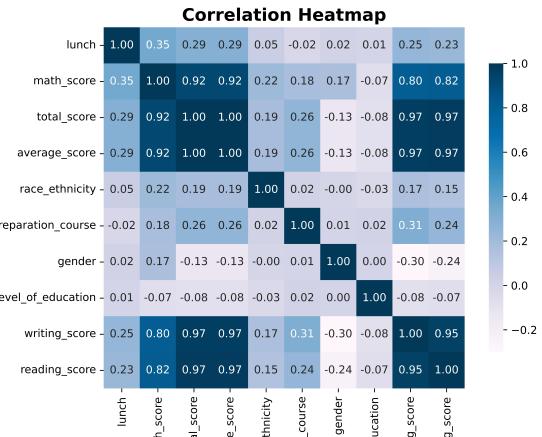


Figure 3: Correlation Heatmap of Variables.

## 4 Causal Discovery Procedure

In this section, we provide a detailed description of the causal discovery process implemented by Causal Copilot. We also provide the chosen algorithms and hyperparameters, along with the justifications for these selections.

### 4.1 Data Preprocessing

In this initial step, we preprocessed the data and examined its statistical characteristics. This process involved data cleaning, handling missing values, and performing exploratory data analysis to examine variable distributions and inter-variable relationships.

### 4.2 Algorithm Recommendation assisted with LLM

Following preprocessing, we employed a large language model (LLM) to assist in selecting appropriate algorithms for causal discovery based on the statistical characteristics of the dataset and relevant background knowledge. The top three chosen algorithms, listed in order of suitability, are as follows:

- **PC:**
  - **Description:** The Peter-Clark Algorithm (PC) is a constraint-based causal discovery method that is flexible in handling both linear and non-linear relationships. It is efficient on CPU and provides strong empirical performance, outputting a CPDAG.
  - **Justification:** The PC algorithm is suitable because it is a constraint-based method that handles both linear and non-linear relationships effectively. It is efficient on CPU and provides strong empirical performance. The algorithm outputs a CPDAG, which is acceptable for the user. It is also scalable to medium-to-large-scale datasets, making it a good fit for the current dataset size.
- **GES:**
  - **Description:** The Greedy Equivalence Search (GES) is a score-based causal discovery method optimized for linear relationships. It is efficient on CPU and provides strong empirical performance, outputting a CPDAG.
  - **Justification:** The GES algorithm is a score-based method optimized for linear relationships but is flexible with noise types. It provides strong empirical performance and is efficient on CPU. Although

it is primarily linear, its strong performance and efficiency make it a suitable choice for this dataset, especially given the homogeneous nature of the data.

Considering data properties, algorithm capability and user's instruction, the final algorithm we choose is PC.

### 4.3 Hyperparameter Values Proposal assisted with LLM

Once the algorithms were selected, the LLM aided in proposing hyperparameters for the chosen algorithm, which are specified below:

- **Significance Level:**
  - **Value:** 0.05
  - **Explanation:** Using a significance level of 0.05 is a standard choice for moderate sample sizes, ensuring a balance between detecting true causal relationships and avoiding false positives.
- **Independence Test Method:**
  - **Value:** fisherz\_cpu
  - **Explanation:** The 'fisherz\_cpu' test is suitable for linear relationships and mixed data types, providing a good balance of accuracy and computational efficiency on CPU.
- **Maximum Depth for Skeleton Search:**
  - **Value:** 5
  - **Explanation:** A depth of 5 is appropriate for a moderate graph size, allowing for comprehensive causal discovery while maintaining computational efficiency.

### 4.4 Graph Tuning with Bootstrap and LLM Suggestion

In the final step, we performed graph tuning with suggestions provided by the Bootstrap and LLM.

We first applied the Bootstrap method to estimate the confidence level associated with each edge in the initial graph. Specifically:

- If an edge not present in the initial graph exhibited a Bootstrap confidence greater than 90%, we added it to the graph.
- Conversely, if an existing edge had a confidence lower than 10%, we removed it.
- For edges with moderate confidence (between 10% and 90%), we consulted the LLM to assess their validity and directionality, drawing on its extensive background knowledge.

The LLM contributed by:

- Reintroducing plausible edges that may have been overlooked by statistical methods;
- Removing or redirecting edges that appeared statistically valid but were conceptually implausible.

To improve the robustness of LLM-generated suggestions, we employed a voting mechanism. Importantly, LLM recommendations were not allowed to override high-confidence decisions made by the Bootstrap procedure. By integrating insights from both of Bootsrap and LLM to refine the causal graph, we can achieve improvements in graph's accuracy and robustness.

## 5 Causal Graph Summary

### 5.1 Causal Graph Discovered by the Algorithm

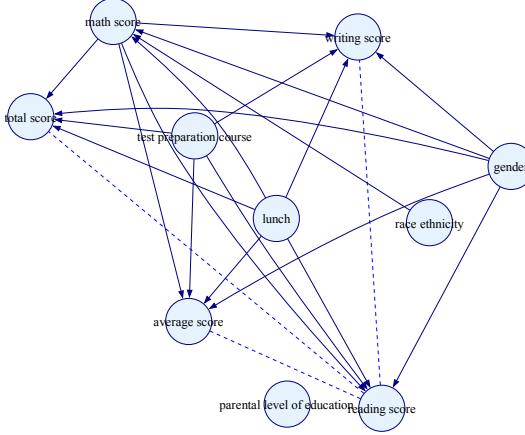


Figure 4: Causal Graph Discovered by the Algorithm. Solid lines represent causal edges identified by the algorithm, while dashed lines indicate strong correlations without inferred causality.

The above is the original causal graph produced by our algorithm.

The causal relationship among the variables suggests a complex interplay where individual characteristics and educational factors influence academic performance across various subjects. Key findings indicate that demographics such as gender and race ethnicity are significant predictors of scores in math, reading, writing, and overall academic performance. Key assessments reveal that parental educational background, reflected through variables such as lunch and test preparation course, plays a role in shaping the academic outcomes. Overall, the analysis highlights that both intrinsic factors (like gender) and external factors (like lunch and test preparation courses) significantly contribute to not just raw scores, but also cumulative measures such as total score and average score.

- Gender directly affects all academic scores, suggesting that there may be systematic differences in performance based on gender, potentially influenced by socio-cultural factors.
- Race ethnicity also impacts math score, which may indicate varying levels of access and resources among differing racial groups.
- Lunch serves as a proxy for socioeconomic status, affecting math, reading, and writing scores, reinforcing the idea that economic factors influence educational outcomes.
- The test preparation course shows a direct causal relationship with reading and writing scores, indicating that preparatory resources can enhance academic performance.
- Math score is a foundational variable that influences both reading and writing scores, which illustrates the interconnectedness of skills in educational achievement.
- Total score and average score are further influenced by multiple factors, emphasizing the need to consider a comprehensive view of student performance that includes both demographic and educational variables.

In conclusion, the analysis reveals that multiple factors, both social and educational, converge to impact students' academic success, highlighting the importance of addressing disparities to improve learning outcomes across diverse populations.

## 5.2 Revised Graph

### 5.2.1 Bootstrap Probability

To evaluate how much confidence we have on each edge, we conducted bootstrapping to calculate the probability of existence for each edge. From the statistics perspective, we can categorize the edges' probability of existence into three types:

- **High Confidence Edges:** gender causes math\_score, lunch causes math\_score, gender causes writing\_score, lunch causes writing\_score, test\_preparation\_course causes writing\_score, math\_score causes writing\_score.
- **Moderate Confidence Edges:** race\_ethnicity causes math\_score, gender causes reading\_score, lunch causes reading\_score, test\_preparation\_course causes reading\_score.
- **Low Confidence Edges:** gender causes total\_score, lunch causes total\_score, test\_preparation\_course causes total\_score, math\_score causes total\_score, gender causes average\_score, lunch causes average\_score, test\_preparation\_course causes average\_score, math\_score causes average\_score.

### 5.2.2 LLM Pruning

By using the method mentioned in the Section 4.4, we provide a revised graph pruned with Bootstrap and LLM suggestion. Pruning results are as follows.

Bootstrap doesn't force or forbid any edges.

The following relationships are forbidden by LLM:

- **gender → writing score:** Gender does not inherently influence writing ability or performance, as writing skills are typically shaped by educational background, personal interests, and practice rather than gender itself;
- **gender → reading score:** Similar to writing, reading skills are developed through education and exposure, so gender does not directly cause changes in reading scores;
- **gender → math score:** Math ability is largely influenced by instructional quality and individual student effort, making gender an irrelevant factor in determining math scores.
- **race ethnicity → math score:** There is no inherent causal relationship between a person's race or ethnicity and their math score, as academic performance in mathematics can be influenced by various factors such as socio-economic status, access to educational resources, and individual effort.
- **race ethnicity → reading score:** Similarly, race or ethnicity does not directly cause variations in reading scores, and any observed differences are more likely attributed to external socio-economic and educational factors rather than to the race or ethnicity itself.

The following are directions confirmed by the LLM:

- **parental level of education → test preparation course:** Higher levels of parental education often lead to greater awareness of educational resources, resulting in increased enrollment in test preparation courses for their children;
- **parental level of education → writing score:** A more educated parent typically emphasizes the importance of literacy and writing skills, which can result in children achieving higher writing scores.
- **lunch → writing score:** School lunch policies can provide students with balanced nutrition, which is essential for cognitive functions and can enhance performance in tasks like writing;
- **lunch → math score:** Access to nutritious school lunches can positively influence students' overall health and energy levels, thereby supporting better concentration and performance in subjects like math.
- **test preparation course → reading score:** Enrolling in a test preparation course typically provides students with strategies and practice that enhance their reading skills, leading to improved performance on reading assessments;
- **test preparation course → writing score:** Participation in a test preparation course often involves writing exercises and feedback, which can significantly enhance students' writing skills, resulting in higher writing scores.

- **math score → writing score:** A higher math score typically reflects better analytical and problem-solving skills, which can enhance a student's ability to articulate and structure their writing, resulting in improved writing scores;
- **math score → reading score:** Proficiency in math can enhance cognitive skills such as logical reasoning and comprehension, which are also essential for strong reading abilities, suggesting that improvements in math scores could lead to better reading scores.
- **reading score → writing score:** Strong reading skills promote better writing abilities, as both involve comprehension and expression of language, hence enhancing a student's overall literacy performance;
- **total score → average score:** The total score, being the sum of individual scores, directly influences the average score, as the average is calculated by dividing the total score by the number of subjects assessed.

This structured approach ensures a comprehensive and methodical analysis of the causal relationships within the dataset.

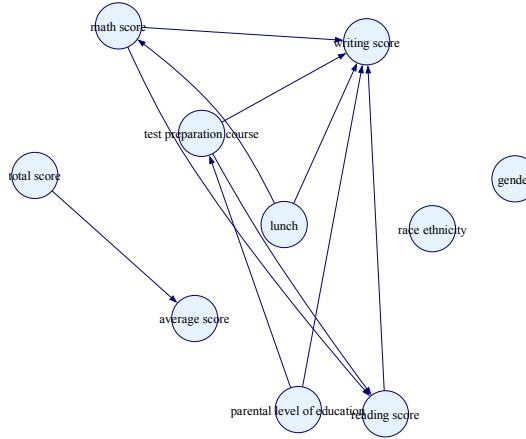


Figure 5: Revised Graph by LLM.

### 5.3 Graph Reliability Analysis

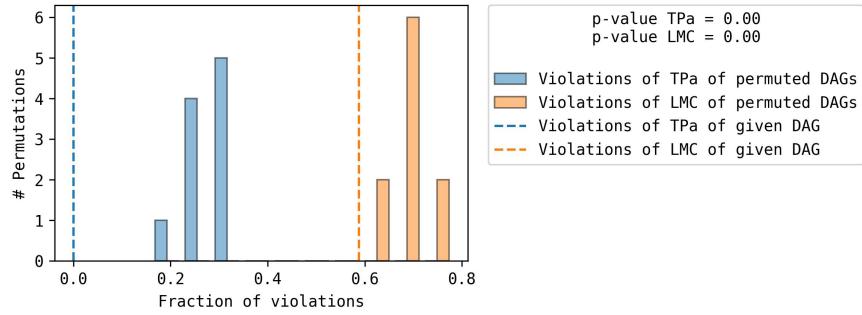


Figure 6: Refutation Graph.

The results from the causal graph refutation test indicate that the provided Directed Acyclic Graph (DAG) is both informative and reliable, as evidenced by the test's outcomes. With 0 out of 10 permutations falling within the Markov equivalence class of the DAG and a significant p-value of 0.00, it suggests that the graph is distinctly characterized and not a mere random configuration. Moreover, the fact that the DAG violates 40 out of 68 linear Markov conditions (LMCs) yet performs better than 100.0% of the permuted DAGs further reinforces its robustness, indicating that the level of violations is not statistically uncharacteristic. Given that both tests' results align with the significance level

of 0.05 — specifically, the DAG’s contextual performance among permuted graphs and its informative nature — we conclude that we do not reject the DAG. Thus, this analysis affirms the reliability of the causal relationships posited by the graph, supporting its validity in causal inference studies.

#### 5.4 Result Graph Comparision

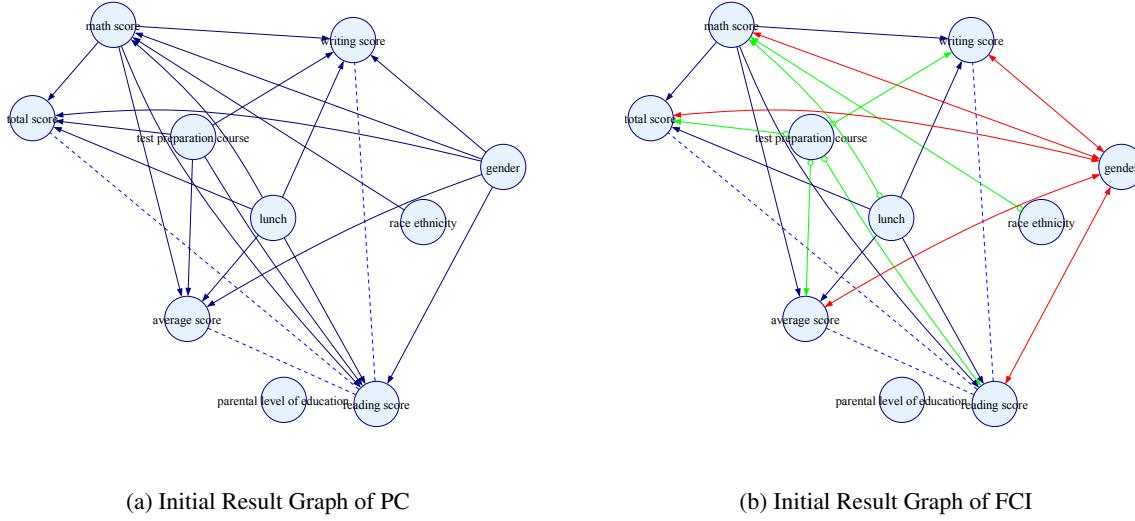


Figure 7: Result Graph Comparision of Different Algorithms.

The causal graphs derived from the PC and FCI algorithms exhibit both unique and overlapping directed edges among variables. The PC algorithm identifies a broader range of directed edges, including causal relationships from gender, race\_ethnicity, and lunch to various outcome measures such as math\_score, reading\_score, writing\_score, total\_score, and average\_score. In contrast, the FCI algorithm focuses on a more limited set of directed edges, primarily involving lunch and math\_score impacting reading\_score, writing\_score, total\_score, and average\_score.

Both algorithms agree on several edges, specifically the relations where lunch and math\_score have a causal influence on reading and writing scores, as well as their impact on total\_score and average\_score. However, the PC algorithm includes additional edges, suggesting a more complex causal structure involving gender and race\_ethnicity that the FCI algorithm does not fully capture due to its focus on hidden confounders.

Edges identified by the PC algorithm may be considered more reliable regarding the overall causal structure since it provides a more comprehensive view of the relationships among variables, including the contributions of gender and race\_ethnicity, which are important contextual factors. The FCI algorithm, while it acknowledges the presence of hidden confounders related to gender, presents fewer direct causal edges, possibly limiting its applicability in scenarios where understanding the full causal network is crucial. Hence, edges supported by both algorithms, particularly those involving lunch and math\_score, may be deemed more robust due to their repeated verification across different methodologies.

#### 5.5 Conclusion

In this causal discovery report, we analyzed a dataset concerning student performance, which included demographic attributes such as gender and race/ethnicity, socioeconomic indicators like lunch type, and academic performance metrics across math, reading, and writing. Our methodology employed the Causal Copilot for data preprocessing, algorithm selection through a large language model (LLM), and subsequent graph tuning utilizing Bootstrap methods, resulting in a refined causal graph illustrating the intricate relationships among these variables.

The results indicated that both intrinsic factors, like gender, and external factors, such as lunch type and educational resources, significantly influence student performance. Notably, we identified that access to lunch positively impacts average scores, with an estimated increase of about 0.92 points. This study contributes to the field by delineating the complex interplay of various demographic and educational factors influencing academic outcomes, providing

educators and policymakers with actionable insights to address disparities in student performance and enhance learning experiences through targeted interventions.

## 6 Causal Inference Results

### 6.1 Proposal Overview

The proposed task aims to investigate the causal relationship between lunch options and students' average scores, focusing on estimating the average treatment effect (ATE) of different lunch types. This inquiry is particularly relevant in the context of educational outcomes, where understanding the factors that influence academic performance is crucial. By analyzing how various lunch options affect students' scores, we can provide valuable insights into the role of nutrition and meal variety on academic achievement. This is significant for educators and policymakers as it may lead to informed decisions regarding school lunch programs that could enhance student performance.

Utilizing a causal inference framework allows us to rigorously assess the impact of lunch choices, treating them as an intervention that may enhance or hinder academic success. By employing appropriate statistical methodologies, such as propensity score matching or regression analysis, we can control for confounding variables and derive a clearer understanding of the causal effect of lunch on academic outcomes. This evidence-based approach is essential for identifying effective strategies to optimize student performance through dietary interventions, making it a pressing and meaningful task in the realm of educational research.

### 6.2 Treatment Effect Estimation Results

#### 6.2.1 Basic Information Overview

- **Treatment Variable:** lunch.
- **Outcome Variable:** average score.
- **Confounders:** math score, total score, test preparation course, writing score, reading score.
- **Heterogeneous Variables:** gender, race, ethnicity, parental level of education, test preparation course, math score.

#### 6.2.2 Estimation Method & Justification

**Chosen Method:** propensity score

##### Justification

When the number of confounders is small, exact or approximate matching is effective in balancing covariates. PSM estimates treatment probability and matches treated and control units. It's better for continuous variables. By controlling confounders with PSM, we can make causal inference more credible.

#### 6.2.3 Estimation Results

##### Balance Check

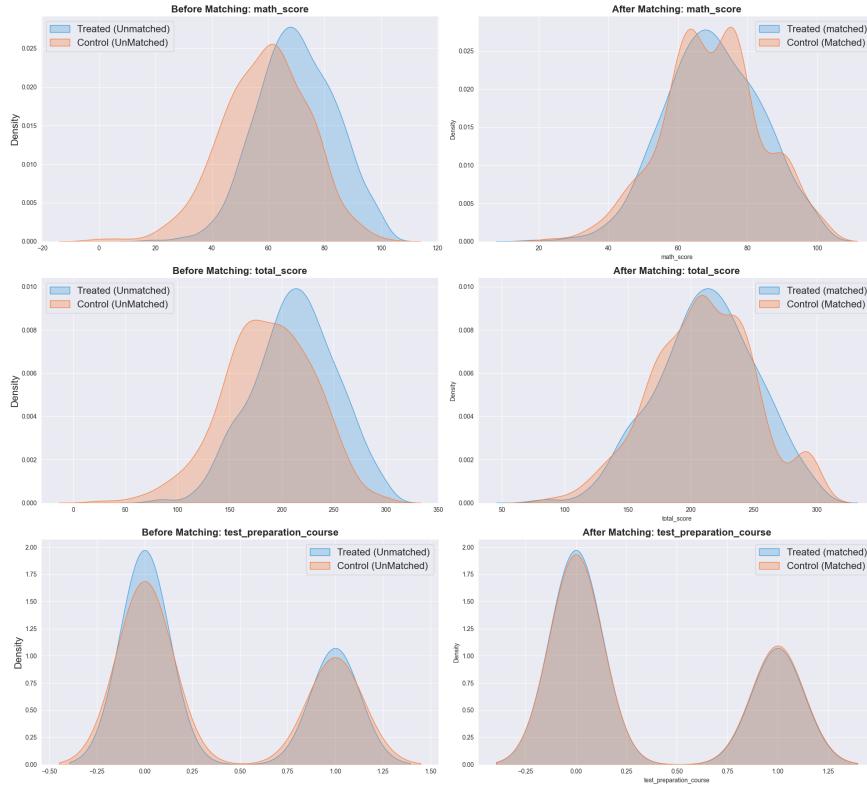


Figure 8: Distribution of Confounders before and after matching.

The figure above is for balance checking. It's a comparison of heterogeneous variables before and after matching.

#### Average Treatment Effect (ATE)

In the analysis of the effect of lunch on average score through matching-based effect treatment estimation, the following results emerged:

- The Average Treatment Effect (ATE) indicates a significant positive impact of lunch on students' average scores, with an estimated increase of approximately 0.92 points for those who received the treatment.
- This effect underscores the potential benefits of providing lunch as a means to enhance academic performance, suggesting that access to lunch may contribute to better engagement and improved outcomes in educational settings.
- The matching method applied effectively controlled for confounding variables, lending credibility to the observed treatment effect and suggesting that the results are robust against bias from those confounders.
- The findings can inform policy discussions regarding the implementation of lunch programs, emphasizing their role in supporting students' academic success.

Overall, the analysis provides strong evidence that lunch has a beneficial effect on student average scores.

### 6.2.4 Conditional Average Treatment Effect (CATE)

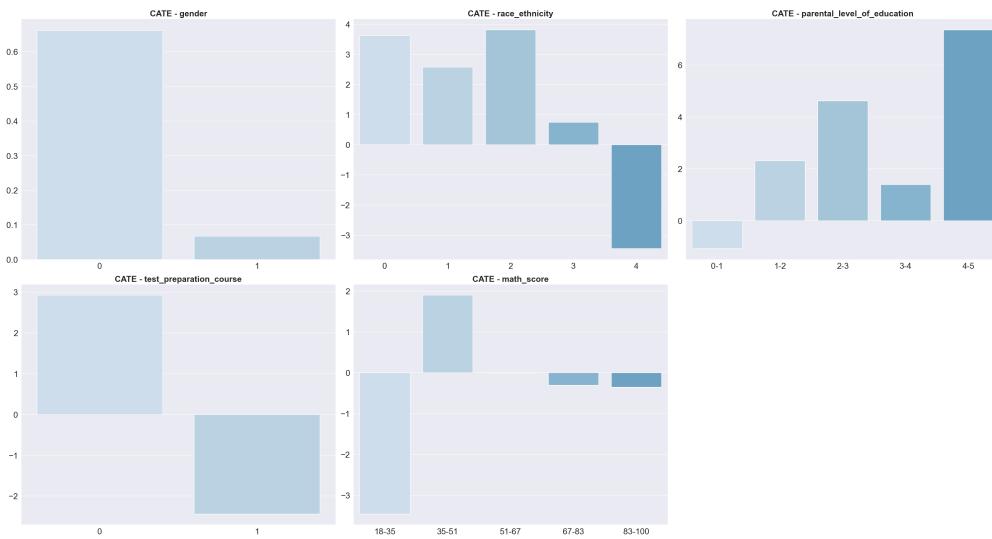


Figure 9: CATE Bar Plots grouped by different confounders.

The analysis of the treatment effect of lunch on average scores, utilizing propensity score matching and accounting for various confounders, reveals notable insights:

- **Overall Impact:** The treatment effect of lunch on average score is significant, with variations based on different conditioning factors.
- **Gender:**
  - Males exhibit a stronger positive treatment effect (+0.66) compared to females (+0.07), suggesting that the impact of lunch might be more beneficial for male students.
- **Race/Ethnicity:**
  - The treatment effect varies considerably across different racial and ethnic groups, with the highest positive impact observed in certain groups, notably +3.61 for one group, contrasting with a negative treatment effect for another (-3.45).
- **Parental Level of Education:**
  - Students with lower parental education levels show a negative treatment effect (-1.07), while those with higher education attainments benefit more, with the highest effect observed at +7.35 for the highest parental education level. This could indicate that the benefits of lunch are amplified with increased parental educational background.
- **Test Preparation Course:**
  - Participation in a test preparation course shows mixed effects: a positive effect (+2.92) for those who took the course versus a negative effect (-2.44) for those who did not, suggesting variability based on preparation resources.
- **Math Scores:**
  - The treatment effect in relation to math scores is mixed, with negative effects among lower scoring ranges (-3.45 at the lowest range) contrasted with marginally positive to neutral effects across mid to high scoring students, indicating that the impact of lunch may differ based on students' prior academic performance.

In summary, the treatment effect of lunch on average scores is nuanced, influenced by gender, racial/ethnic backgrounds, parental education, preparation courses, and student performance. This suggests that interventions aimed at enhancing lunch programs might benefit from customization based on these demographic and academic factors.

## 6.3 Summary & Next Steps

### 6.3.1 Discussion

The findings indicate a **positive relationship** between having lunch and the average scores of students, with an Average Treatment Effect (ATE) revealing an estimated increase of approximately **0.92 points for those who had lunch**. This result strongly suggests that access to lunch can serve as a significant factor in enhancing academic performance, ultimately contributing to improved educational outcomes and student engagement.

Furthermore, the analysis uncovers **variability in the treatment effect across different demographic factors** such as gender, race/ethnicity, parental education level, and participation in test preparation courses. This highlights the importance of considering these variables, as they may influence the degree to which lunch contributes to academic success. Such insights underscore the need for targeted interventions that accommodate diverse student backgrounds to effectively leverage the benefits of lunch in educational settings.

### 6.3.2 Next Steps Suggestions

To enhance the robustness of our findings, **some potential improvements include performing sensitivity analyses to further explore the impact of unobserved confounders and validating the propensity score matching results through alternate methods such as inverse probability weighting**. Additionally, expanding the analysis to include longitudinal data could help us better understand the sustained effects of lunch programs over time. Future research should aim to incorporate a larger and more diverse sample to validate these findings across different educational settings and demographics. **Exploring qualitative dimensions, such as students' perceptions of the lunch program and its motivational effects, can also provide deeper insights into why these variations occur**.

Moreover, to expand upon the insights gained from this study, future investigations could examine the long-term academic trajectories of students benefiting from lunch programs and encompass broader outcomes beyond average scores, such as social and emotional well-being. **Such an approach can reveal the multifaceted impacts of lunch provision in schools, informing more comprehensive policy frameworks**. Collaborating with educational policy-makers and practitioners would ensure that findings are translated into actionable strategies, and longitudinal studies could assess how changes in program implementation might influence academic and non-academic outcomes over time.