Final Report

2024-12-08

Final report

- Group Topic: Health
- Team members: Mahitha Penmetsa (msp259), Ram Peddu (rbp94), Hermon Beluts (htb35), Isabella Censullo (ilc8), Devansh Khadka (dbk84)

Causal Question

Describe your causal question in a way that someone who has not taken this class would understand. Why are you interested in this question? How could answering this question allow for better decision making? Include any necessary background or context. Cite outside sources you use. Answer

The causal question we are examining is "Does the quality of sleep affect a students academic performance as measured by their exam scores?" We would like to establish if sleeping more hours, specifically defining good sleep as anything over 6 hours, will positively influence the student's exam performance. We all have discussed how we sleep very little at school and how that may make it difficult for us to study. Answering this question could allow for better decision-making by establishing a clear link between better sleep and better academic performance, allowing schools and families to adjust their scheduling/workload to accommodate students' needs. Students often have to balance work, homework, extracurriculars, etc., preventing them from establishing quality sleep hygiene. However, leveraging quality sleep to optimize student performance could resolve this.

Describe your causal question in the language of causal inference we've learned in this course: What is the treatment? What is the outcome? What are the potential outcomes? Write these out in words and in the math notation we have used in class. Answer

The treatment is the number of hours of sleep a student receives per night, and the outcome is their final exam score.

Potential Outcomes Notation

Let A_i represent the binary treatment variable where:

$$A_i = \begin{cases} 0 & \text{if Sleep Hours } \leq 6 \\ 1 & \text{if Sleep Hours } > 6 \end{cases}$$

This represents whether students received above or below an adequate 6 hours of sleep.

Let Y_i represent the outcome of a particular student i, where the potential outcomes are as follows:

- Y_i^0: Represents the final exam score for student i if they get 6 or less hours of sleep per night ($A_i = 0$).
- Y_i^1: Represents the final exam score for student i if they get more than 6 hours of sleep per night ($A_i = 1$).

Causal Diagram

Draw a DAG representing your causal question that includes at least three relevant variables besides treatment and outcome that are included in your dataset. You may include more than three variables. You may include variables that are not in your dataset, but at least 3 of your variables (excluding treatment and outcome) must be included in your dataset. If you use letters to denote variables, make sure they are clearly defined Answer

Figure 1: DAG

- E-> Participates in extracurricular activities
- R-> Availability of educational resources
- L-> Presence of learning disabilities
- Q-> Teacher Quality
- T-> Tutoring Sessions attended or not

Explain your DAG: tell us in words what is meant by each edge in your DAG. Answer

1. Extracurricular Activities (E) $\rightarrow A_i$: Participation in extracurricular activities might reduce the number of hours available for sleep due to time constraints. Students who are highly engaged in these activities often sacrifice sleep to manage their schedules.
2. Educational Resources (R) $\rightarrow Y_i$: Access to educational resources, such as tutoring, study materials, or a conducive learning environment, directly influences academic performance. Students with more resources may perform better on exams, regardless of sleep quality.
3. Learning Disabilities (L) $\rightarrow Y_i$: Learning disabilities can affect a student's ability to grasp material and perform on exams, independent of sleep quality. This serves as an important confounder in our analysis.
4. Teacher Quality (Q) $\rightarrow Y_i$: The quality of instruction a student receives can heavily impact their exam performance. A high-quality teacher might improve learning outcomes, thus influencing exam scores.
5. Tutoring Sessions (T) $\rightarrow A_i$: Students attending tutoring sessions may have better academic support, but this could also lead to reduced sleep hours if tutoring occurs late at night or adds to their workload.

6. A_i $\rightarrow Y_i$: The direct causal link of interest. Here, we hypothesize that adequate sleep improves academic performance.
7. Extracurricular Activities (E) $\rightarrow Y_i$: Participation in extracurriculars could have a dual effect: while it may improve non-cognitive skills beneficial to academic performance, it could also reduce the time available for studying and sleeping.
8. Learning Disabilities (L) $\rightarrow A_i$: Students with learning disabilities may face more academic pressure, leading to less sleep due to the time required for studying or completing assignments.
9. Tutoring Sessions (T) $\rightarrow Y_i$: Having access to tutoring sessions will likely help students prepare for exams and probably will help them achieve higher exam scores.

Discuss your DAG. How realistic is it? Are there variables or edges you excluded from your DAG that someone else might argue should be included? Playing devil's advocate, how would you critique the reliability of your DAG? Answer

Our DAG attempts to capture the most critical relationships influencing the causal link between sleep quality (A_i) and academic performance (Y_i), and it is reasonably realistic. It includes plausible variables like extracurricular activities (E), which impact time management and sleep duration, as well as educational resources (R), teacher quality (T), and learning disabilities (L), all of which are well-documented contributors to academic outcomes. However, there are potential exclusions worth considering. Mental health, for example, could influence both sleep quality and academic performance through factors like anxiety or depression, creating bidirectional relationships that the DAG does not address. Similarly, parental involvement may affect students' study schedules and sleep hygiene, and socioeconomic status (SES) could influence access to resources, teacher quality, and the ability to establish regular sleep routines. Excluding these variables could lead to omitted variable bias. The DAG assumes linear and unidirectional relationships, which may oversimplify real-world dynamics, such as feedback loops between stress and academic performance. Additionally, measurement error in key variables like sleep quality and exam scores could bias the results. A critic might argue that the DAG oversimplifies the multifaceted nature of sleep and academic performance by excluding psychological, physiological, and social factors or by focusing on a binary threshold of "over 6 hours" without accounting for nuances in sleep quality. Despite these limitations, the DAG is a realistic and practical representation for this analysis, but careful interpretation is needed to address the potential for unmeasured confounding and simplifying assumptions.

Method and Identification

What method are you using to estimate a causal effect? What causal effect are you estimating (ATE vs LATE vs ATT)? What assumptions are required to identify the causal effect via your chosen method? Answer

We are using Propensity Score Matching to estimate the causal effect. This means we match people based on their likelihood of receiving the treatment to control for confounding variables. We are estimating the average treatment effect on the treated (ATT), because we are comparing the difference in school performance results of the individuals in the control vs treatment group. The one assumption we rely on to identify the causal effect is conditional exchangeability, i.e that given a set of confounders, the potential outcome is independent of the treatment.

Explain what conditional exchangeability means in the context of your causal question. Is it important? Why or why not? How do sufficient adjustment sets relate to conditional exchangeability? Answer

Conditional Exchangeability means that after adjusting for observed confounders, exam scores would only be affected by the hours of sleep a person has had. Essentially, the distribution of exam scores under adequate and inadequate sleep conditions should be similar to if the hours of sleep were randomly sampled. It is extremely important because without conditional exchangeability, unaccounted confounders could bias the estimated effect of sleep on exam scores. A sufficient adjustment set is a set of covariates that blocks all

non-causal paths between sleep and exam performance in the DAG. If the adjustment set has all possible confounders then conditional exchangeability holds.

Assuming your DAG is true, list out all non-causal paths between treatment and outcome and list one sufficient adjustment set to identify the causal effect of the treatment on the outcome. If a sufficient adjustment set does not exist, add additional variables to your DAG so that one does exist. Answer

Here is the list of non causal paths: Sleep <- Extracurricular Activities -> Academic Performance, Sleep <- Tutoring Sessions -> Academic Performance, Sleep <- Learning Disability -> Academic Performance. Therefore, one sufficient adjustment set to identify the causal effect would be : { Extracurricular Activities, Tutoring Sessions, Learning Disability}

Discuss the plausibility of conditional exchangeability in your setting. If your sufficient adjustment set contains variables that are not in your dataset, discuss the implications. Answer

While our DAG does cover many major confounders, like learning disabilities and extracurricular activities, that would reasonably affect the outcome if not blocked, it does also miss other potential factors. Some of these factors could be mental health and socioeconomic conditions, which are not in our adjustment set but could theoretically impact both the treatment and outcome. This leads to a potential implication that our results may be biased and either overestimate or underestimate the causal effect.

Discuss any other identification assumptions for your method here, such as positivity and consistency. What do they mean in the context of your causal question and are they plausible? Answer

For positivity to hold, the students in our dataset would have to have a non-zero chance of receiving either adequate or inadequate sleep, which would seem to be true as one can likely control the amount of sleep they receive. For consistency to hold, the treatment has to be well-defined and the observed outcomes would have to match the defined treatment levels. Tis is true in our case as we clearly state that adequate sleep corresponds to six or more hours.

Discussion: Analysis and Results

Give some context for your dataset. Who is included in your dataset? How was the data collected? When was the data collected? Make sure to cite the dataset. Answer

The Kaggle dataset explores various factors affecting student performance in exams, including aspects like attendance, parental involvement, access to resources, and more. The data was conducted on 6607 students and includes 20 features. Based on the nature of the features, it appears that the data was collected through either observational study or self-reported surveying. While the time period of data collection isn't specifically stated, the dataset was uploaded to Kaggle in 2023. (https://www.kaggle.com/datasets/lainguyn123/student-performance-factors/data)

Discuss any choices you made regarding data cleaning and processing: Did your data have missing values or outliers? How did you handle them? Were there any variables you dichotomized (i.e. made binary), or variables that you changed the format (e.g. yes/no to 1/0)? Answer

We began the data cleaning process by exploring the dataset for missing or NA values. Missing values appear for variables TeacherQuality, DistancefromHome, and FamilyIncome, but since these features weren't relevant to our analysis we chose not to address them. Using the summary() function, we were able to determine that there weren't any missing values in our treatment and outcome variables. We proceeded to explore SleepHours for potential outliers or irregularities in the data using a boxplot, and it appeared to be clean (reasonably spread data from 4-10 without outliers). Following this, we explored ExamScores to ensure that they ranged up to 100, logically. We used the str() function to confirm that the dataset was clean and

well-structured. This was the extent of our data cleaning process as the dataset was already fairly clean beforehand. In terms of processing, we converted the continuous variable SleepHours into a binary treatment variable of either adequate or inadequate sleep. The ifelse() function was used to create a new variable called treatment and convert it to a factor. We also converted several other confounding variables to factors. We scaled ExamScores to be standardized in order to put variables on the same scale to interpret more easily.

Discuss the impact of any choices you made regarding your dataset, such as choices you made in data cleaning or processing. Answer

The impact of converting our treatment to a binary treatment variable is that it allows for a more simple comparison between treatment and control groups. It also allows for the application of matching techniques further in our analysis. The impact of factor conversion is that it helps when using categorical variables to provide the correct structure for interpretation and during matching analysis.

Explain how you estimated a causal effect.

- If you used matching, explain and discuss your choices. What formula did you use and why? What matching strategy did you use and why? Are there any advantages or drawbacks to the strategy you chose? How many units did your matching drop? How was the covariate balance in your matched sample? Discuss the implications of any choices you made and the quality of your matching.
- If you didn't use matching, explain any choices you made related to the method you used and discuss their implications. Think about advantages or drawbacks to any choices you made, possible bias-variance trade-offs, and assessing how well your method did.

Answer

We used Propensity Score Matching, estimating the average treatment effect on the treated (ATT). The formula we used was $treatment \sim ExtracurricularActivities + TutoringSessions + LearningDisabilities$ because we identified these covariates as part of our sufficient adjustment set. This means that when these covariates are accounted for, this ensures the causal effect of the treatment on the outcome can be unbiasedly estimated by blocking all backdoor paths. We used optimal matching, which minimizes the distance between matched treated and control units. The advantages of optimal matching are that you don't need to specify the order in which units are matched and it minimizes the likelihood that extreme within-pair distances will be large, unlike with nearest neighbor matching. One possible disadvantage is that this method requires more computation for larger datasets than nearest neighbor matching. (Source: https://kosukeimai.github.io/MatchIt/reference/method_optimal.html) From our matching output, 1,847 control units were dropped. The covariate balance in the matched sample was refined (more balanced) based on the standardized mean difference, which should be close to 0, before and after matching. For example, the std. mean diff. for Learning_Disabilities before matching was 0.0361 compared to 0.0000 after. The matching strategy successfully balanced the observed covariates.

Report your causal effect estimate and interpret it in the context of your causal question. Answer

The estimate for the Average Treatment Effect on the Treated (ATT) is -0.0373679. This means that students who received the 'Adequate Sleep' treatment, more than six hours of sleep, scored 0.0374 standard deviations lower on their exams, on average, compared to similar students who received inadequate sleep (6 hours or fewer). However, the estimate is very close to zero which could indicate that it's not very meaningful and the sleep treatment doesn't significantly improve exam scores.

Discuss the limitations of your analysis: what are the limitations of your dataset? Is there other data you would have wanted to have to bolster your analysis? Playing devil's advocate, how would you critique the reliability of you causal estimate? Answer

The limitations of the dataset are that the only information on sleep that we have are SleepHours. This doesn't account for sleep quality, sleep interruption, or the fact that our determination of "adequate" sleep

can vary across individuals. This kind of data may have been helpful in bolstering our analysis. Additionally, we only accounted for three covariates in our matching while the dataset contains many other features that could be impacting academic performance, as well as a number of other factors that weren't accounted for such as stress or socioeconomic status. The reliability of the causal estimate could be questioned due to how small it is and the fact that unmeasured confounders could have affected exam performance.

Code:

```
#### Loading Packages ####
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(readr)
library("tidyverse")
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library("MatchIt")
```

```
#### Exploring data ####
summary(data$Sleep_Hours) # No missing values
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.000   6.000   7.000   7.029   8.000  10.000
```

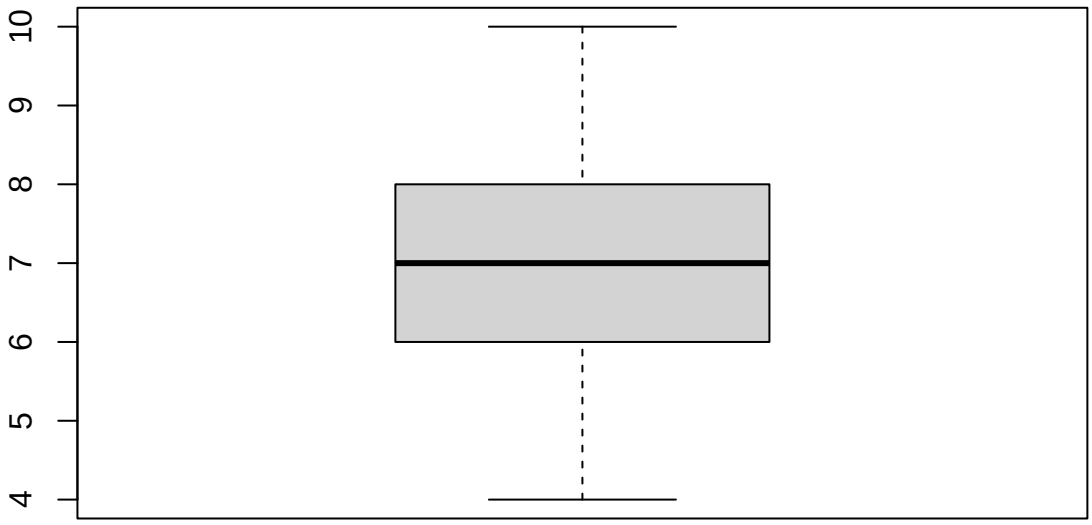
```
summary(data$Exam_Score) # No missing values
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   55.00   65.00   67.00   67.24   69.00  101.00
```

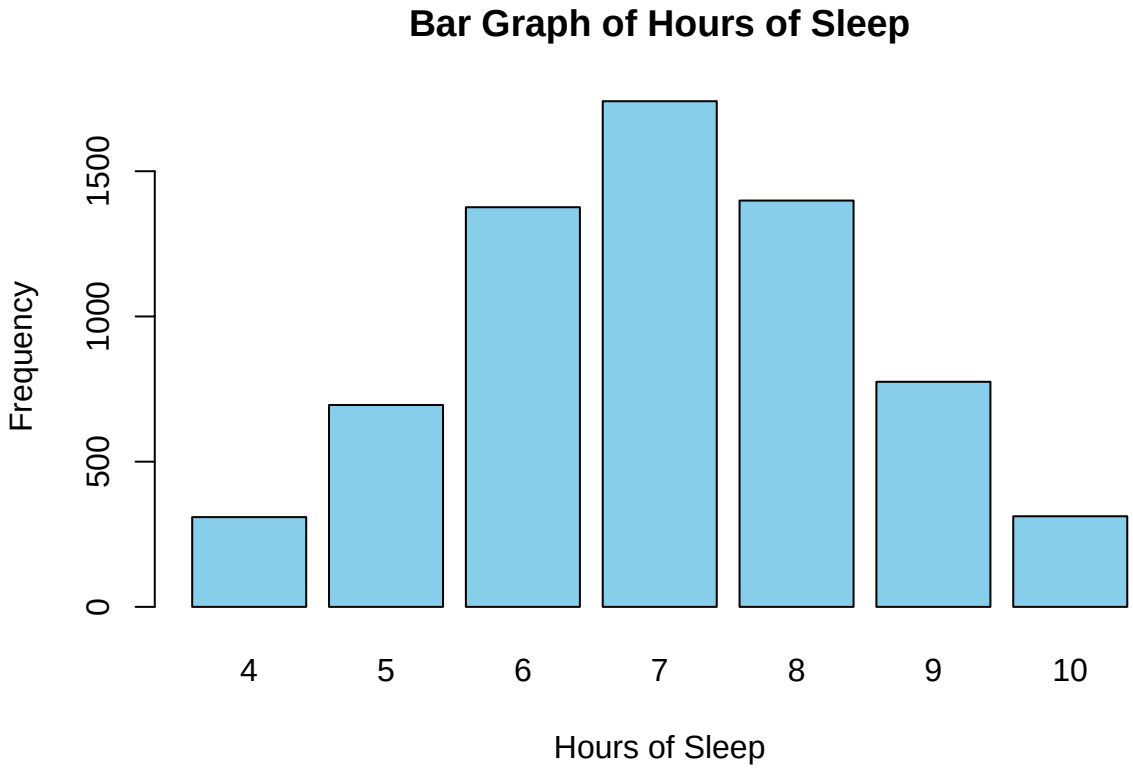
```
anyNA(data) # Missing values appear for variables Teacher_Quality, Distance_from_Home, Family_Income - 
```

```
## [1] TRUE
```

```
boxplot(data$Sleep_Hours) # No outliers
```




```r
sleep_table <- table(data$Sleep_Hours)
barplot(sleep_table,
        main="Bar Graph of Hours of Sleep",
        xlab="Hours of Sleep",
        ylab="Frequency",
        col="skyblue",
        border="black") # create bar graph to visualize hours of sleep distribution in data
```




```r
range(data$Exam_Score)
```

```
## [1]  55 101
```

```r
print(nrow(data))
```

```
## [1] 6607
```

```
colnames(data)
```

```
##  [1] "Hours_Studied"              "Attendance"
##  [3] "Parental_Involvement"       "Access_to_Resources"
##  [5] "Extracurricular_Activities" "Sleep_Hours"
##  [7] "Previous_Scores"            "Motivation_Level"
##  [9] "Internet_Access"            "Tutoring_Sessions"
## [11] "Family_Income"              "Teacher_Quality"
## [13] "School_Type"                "Peer_Influence"
## [15] "Physical_Activity"          "Learning_Disabilities"
## [17] "Parental_Education_Level"   "Distance_from_Home"
## [19] "Gender"                     "Exam_Score"
```

```
str(data)
```

```
## spc_tbl_ [6,607 x 20] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Hours_Studied             : num [1:6607] 23 19 24 29 19 19 29 25 17 23 ...
##  $ Attendance                : num [1:6607] 84 64 98 89 92 88 84 78 94 98 ...
##  $ Parental_Involvement      : chr [1:6607] "Low" "Low" "Medium" "Low" ...
##  $ Access_to_Resources       : chr [1:6607] "High" "Medium" "Medium" "Medium" ...
##  $ Extracurricular_Activities: chr [1:6607] "No" "No" "Yes" "Yes" ...
##  $ Sleep_Hours               : num [1:6607] 7 8 7 8 6 8 7 6 6 8 ...
##  $ Previous_Scores           : num [1:6607] 73 59 91 98 65 89 68 50 80 71 ...
##  $ Motivation_Level          : chr [1:6607] "Low" "Low" "Medium" "Medium" ...
##  $ Internet_Access           : chr [1:6607] "Yes" "Yes" "Yes" "Yes" ...
##  $ Tutoring_Sessions         : num [1:6607] 0 2 2 1 3 3 1 1 0 0 ...
##  $ Family_Income             : chr [1:6607] "Low" "Medium" "Medium" "Medium" ...
##  $ Teacher_Quality           : chr [1:6607] "Medium" "Medium" "Medium" "Medium" ...
##  $ School_Type               : chr [1:6607] "Public" "Public" "Public" "Public" ...
##  $ Peer_Influence            : chr [1:6607] "Positive" "Negative" "Neutral" "Negative" ...
##  $ Physical_Activity         : num [1:6607] 3 4 4 4 4 3 2 2 1 5 ...
##  $ Learning_Disabilities     : chr [1:6607] "No" "No" "No" "No" ...
##  $ Parental_Education_Level  : chr [1:6607] "High School" "College" "Postgraduate" "High School" ...
##  $ Distance_from_Home        : chr [1:6607] "Near" "Moderate" "Near" "Moderate" ...
##  $ Gender                    : chr [1:6607] "Male" "Female" "Male" "Male" ...
##  $ Exam_Score                : num [1:6607] 67 61 74 71 70 71 67 66 69 72 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Hours_Studied = col_double(),
##   ..   Attendance = col_double(),
##   ..   Parental_Involvement = col_character(),
##   ..   Access_to_Resources = col_character(),
##   ..   Extracurricular_Activities = col_character(),
##   ..   Sleep_Hours = col_double(),
##   ..   Previous_Scores = col_double(),
##   ..   Motivation_Level = col_character(),
##   ..   Internet_Access = col_character(),
##   ..   Tutoring_Sessions = col_double(),
##   ..   Family_Income = col_character(),
##   ..   Teacher_Quality = col_character(),
##   ..   School_Type = col_character(),
##   ..   Peer_Influence = col_character(),
##   ..   Physical_Activity = col_double(),
##   ..   Learning_Disabilities = col_character(),
```

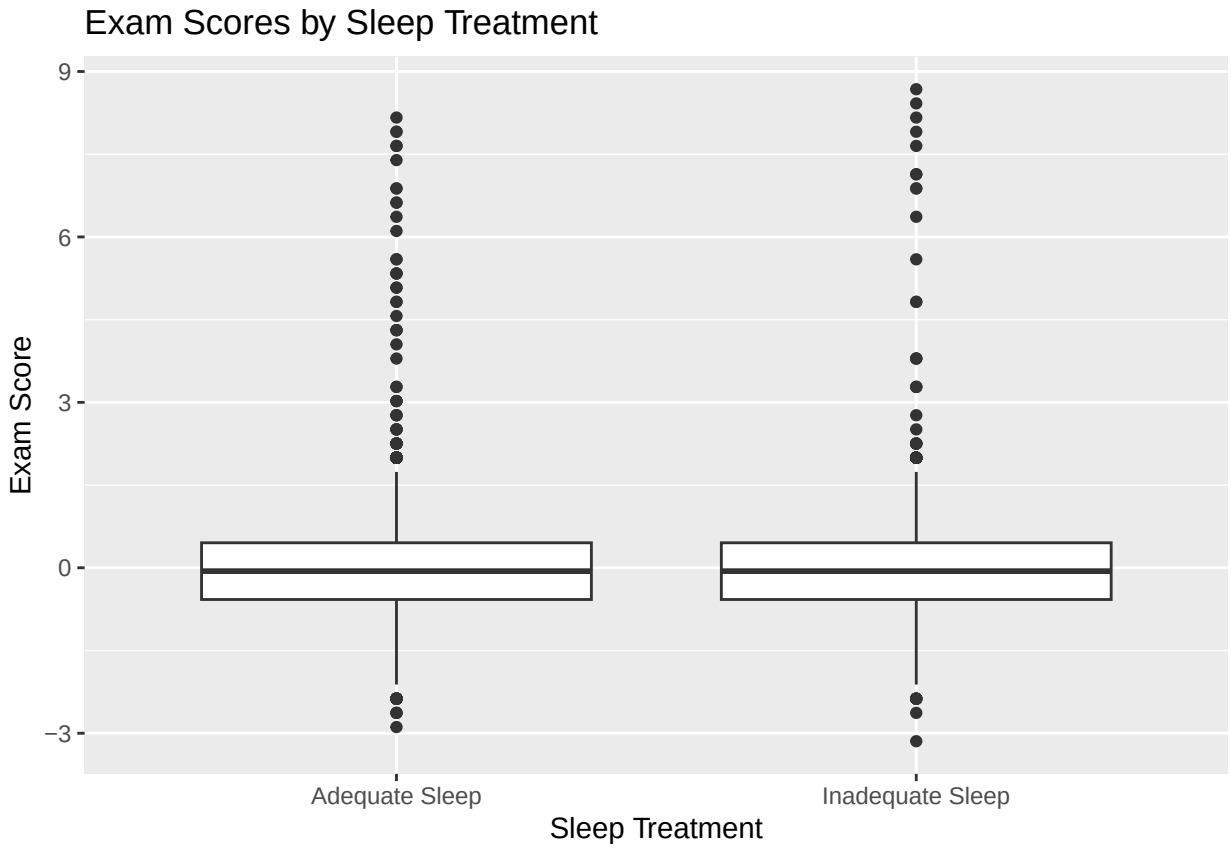
```
##   ..   Parental_Education_Level = col_character(),
##   ..   Distance_from_Home = col_character(),
##   ..   Gender = col_character(),
##   ..   Exam_Score = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
#### Process data ####
# Modify binary treatment variable for adequate sleep (NOTE: The number of hours is meant to be represe
data$treatment <- ifelse(data$Sleep_Hours > 6, "Adequate Sleep", "Inadequate Sleep")
data$treatment <- as.factor(data$treatment)
table(data$treatment)
```

```
##
##   Adequate Sleep Inadequate Sleep
##             4227             2380
```

```
# Standardize exam scores (put variables on the same scale to interpret more easily) - NOTE: this is op
data$Exam_Score <- scale(data$Exam_Score)

# Visualization Example
ggplot(data, aes(x = treatment, y = Exam_Score)) +
  geom_boxplot() +
  labs(title = "Exam Scores by Sleep Treatment", x = "Sleep Treatment", y = "Exam Score")
```




```
# Convert Motivation Level into factors
data$Motivation_Level <- factor(data$Motivation_Level,
                                levels = c("Low",
                                           "Medium",
```

```
                                            "High"),
                                    ordered = TRUE)

# Convert Extracurricular Activities into factors
data$Extracurricular_Activities <- factor(data$Extracurricular_Activities,
                                          levels = c("No", "Yes"),
                                          labels = c("No", "Yes"))

# Convert Internet Access into factors
data$Internet_Access <- factor(data$Internet_Access,
                               levels = c("No", "Yes"),
                               labels = c("No", "Yes"))

# Convert Peer Influence to factors
data$Peer_Influence <- factor(data$Peer_Influence,
                              levels = c("Negative",
                                         "Neutral",
                                         "Positive"),
                              ordered = TRUE)
```

```
#### Optimal Matching ####
matched_data <- matchit(treatment ~ Extracurricular_Activities  + Tutoring_Sessions + Learning_Disabili
summary(matched_data)
```

```
##
## Call:
## matchit(formula = treatment ~ Extracurricular_Activities + Tutoring_Sessions +
##     Learning_Disabilities, data = data, method = "optimal")
##
## Summary of Balance for All Data:
##                               Means Treated Means Control Std. Mean Diff.
## distance                             0.3604        0.3601          0.0395
## Extracurricular_ActivitiesNo         0.4000        0.4062         -0.0127
## Extracurricular_ActivitiesYes        0.6000        0.5938          0.0127
## Tutoring_Sessions                    1.5017        1.4892          0.0099
## Learning_DisabilitiesNo              0.9017        0.8909          0.0361
## Learning_DisabilitiesYes             0.0983        0.1091         -0.0361
##                               Var. Ratio eCDF Mean eCDF Max
## distance                          0.9282    0.0066   0.0239
## Extracurricular_ActivitiesNo           .    0.0062   0.0062
## Extracurricular_ActivitiesYes          .    0.0062   0.0062
## Tutoring_Sessions                 1.0813    0.0031   0.0079
## Learning_DisabilitiesNo                .    0.0107   0.0107
## Learning_DisabilitiesYes               .    0.0107   0.0107
##
## Summary of Balance for Matched Data:
##                               Means Treated Means Control Std. Mean Diff.
## distance                             0.3604        0.3604          0.0006
## Extracurricular_ActivitiesNo         0.4000        0.4013         -0.0026
## Extracurricular_ActivitiesYes        0.6000        0.5987          0.0026
## Tutoring_Sessions                    1.5017        1.5029         -0.0010
## Learning_DisabilitiesNo              0.9017        0.9017          0.0000
## Learning_DisabilitiesYes             0.0983        0.0983          0.0000
##                               Var. Ratio eCDF Mean eCDF Max Std. Pair Dist.
```

```
## distance                          1.0025    0.0004   0.0042          0.0011
## Extracurricular_ActivitiesNo           .    0.0013   0.0013          0.2479
## Extracurricular_ActivitiesYes          .    0.0013   0.0013          0.2479
## Tutoring_Sessions                 1.0488    0.0025   0.0050          0.2915
## Learning_DisabilitiesNo                .    0.0000   0.0000          0.0000
## Learning_DisabilitiesYes               .    0.0000   0.0000          0.0000
##
## Sample Sizes:
##           Control Treated
## All          4227    2380
## Matched      2380    2380
## Unmatched    1847       0
## Discarded       0       0
```

```
#### Estimating ATT ####
only_matched_data <- match.data(matched_data)
ATT <- with(only_matched_data, mean(Exam_Score[treatment == "Adequate Sleep"]) - mean(Exam_Score[treatm
print(ATT)
```

```
## [1] -0.0373679
```

```
```