|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Variable** | **Control (Mean+SD)** | **Treatment (Mean+SD)** | **P values** |
| 1. | Age (in months) | -1.83±69.8 | 0.08±68.8 | 0.694 |
| 2. | Matrix Reasoning | 4.34±2.99 | 4.17±2.92 | 0.419 |
| 3. | Bug Search | 10.41±6.39 | 10.69±6.76 | 0.56 |
| 4. | Picture Memory | 4.21±3.31 | 4.06±3.44 | 0.539 |
| 5. | Vocabulary | 4.03±4.25 | 3.53±4.0 | 0.086 |
| 6. | Animal Coding | 8.31±4.9 | 8.61±5.55 | 0.45 |
| 7. | Receptive Vocabulary | 5.19±4.4 | 5.24±4.4 | 0.86 |
| 8. | Picture naming | 4.27±2.69 | 4.196±2.84 | 0.703 |
| 9. | ASER Math | 0.326±0.38 | 0.265±0.29 | 0.201 |
| 10. | ASER Kannada | 0.293±0.32 | 0.265±0.29 | 0.201 |
| 11. | Motor Skills | 0.000±0.5 | -0.032±0.51 | 0.43 |
| 12. | Body Parts | 4.74±2.69 | 4.30±2.90 | 0.026 |
| 13. | Colors | 1.854±2.00 | 1.903±2.09 | 0.731 |
| 14. | DIAL number | 1.644±1.34 | 1.493±1.39 | 0.118 |
| 15. | Actions | 5.74±4.14 | 5.022±4.27 | **0.015** |
| 16. | Alphabets | 5.72±7.02 | 4.91±6.32 | 0.086 |
| 17. | Problem Solving | 2.33±2.6 | 2.13±2.5 | 0.283 |
| 18. | Copying | 3.30±3.68 | 3.03±3.70 | 0.300 |
| 19. | Panamath | 3.47±2.99 | 2.96±2.93 | **0.015** |
| 20. | Total Score | -1.35±0.73 | -1.38±0.78 | 0.50 |

**Task 1: Table Describing Baseline Balance**

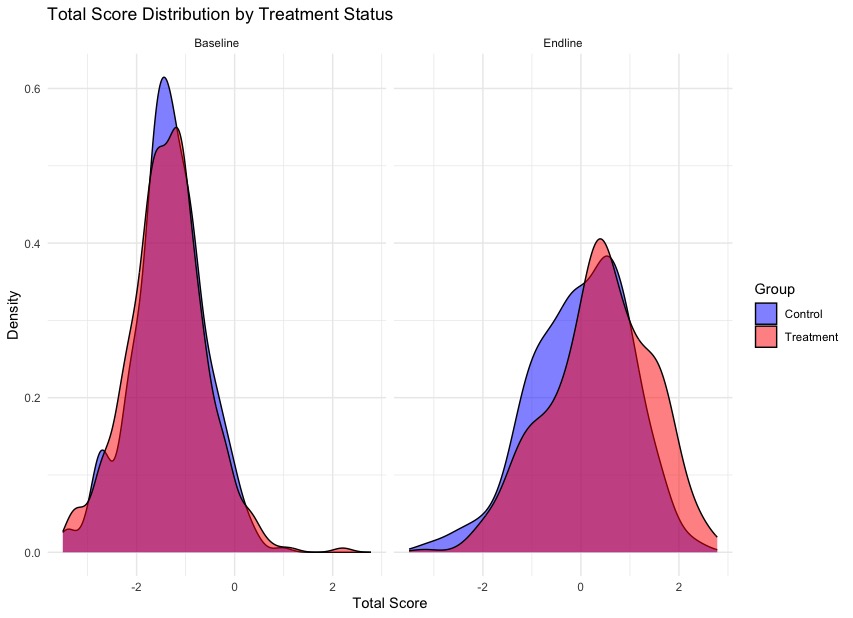
**Task 2:**

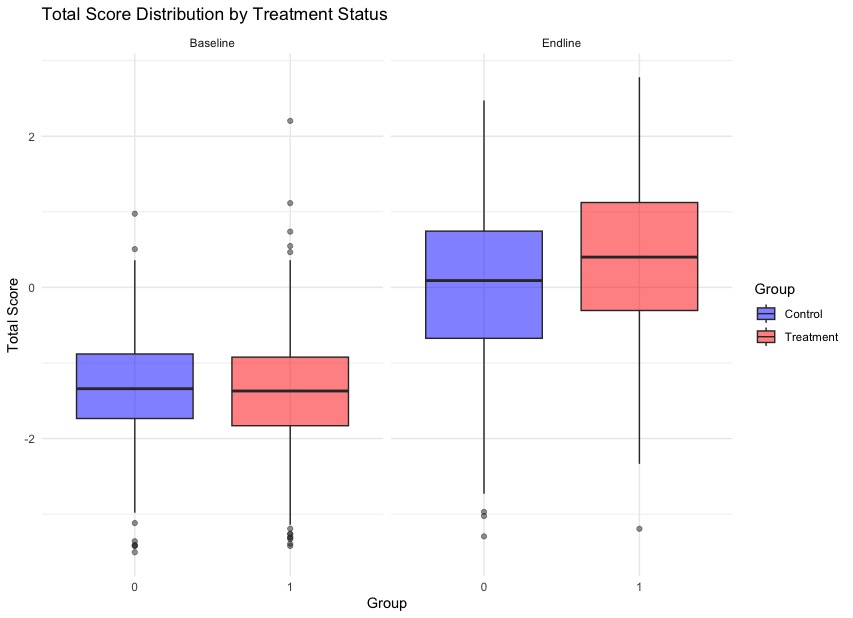
**Percentage of Guardians who disagree on the type of primary school they are planning for children:**

Answer is 7 % (please find the codeset in the R script)

**Task 3:**

**Graphical representation of total score for baseline and endline, bifurcated for control groups and treatment groups is given below as i) scatterplot, ii) boxplot**

****



**Task 4:**

**Table 2: Regression of endline scores against baseline scores, age, mother’s education and treatment vs control group**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Cognitive Index (Endline)** | **Language Index (Endline)** | **Quantitative Index (Endline)** | **Motor Index (Endline)** |
| **Intercept** | 0.499 | 0.940 \* | 0.460 | 0.440 |
| **Baseline Index** | 0.038 | 0.060 | 0.125 | 0.054 |
| **Age** | 0.001 | -0.007 | 0.006 | 0.003 |
| **Education** | -0.009 | -0.003 | -0.011 | -0.007 |
| **Treatment** | -0.020 | -0.020 | 0.005 | -0.012 |
| **R2** | 0.007 | 0.005 | 0.007 | 0.005 |
| **Adjusted R2** | -0.0004 | -0.0032 | -0.0008 | -0.0029 |
| **F-Statistics** | 0.950 (p = 0.43) | 0.595 (p = 0.667) | 0.895 (p = 0.466) | 0.631 (p = 0.641) |
| \*\*\*p < 0.001, \*\*p < 0.01, \*p < 0.05 | | | | |

**Interpretation:**

* There is no significant effect of treatment group, and treatment vs non-treatment groups did not show a statistically significant difference.

**R Code:**

install.packages("purrr")

library(tidyverse)

library(readxl)

library(purrr)

library(ggplot2)

library(writexl)

getwd()

child\_data <- read.csv("/Users/anshulkandpal/Downloads/be\_dataset/childdata.csv")

guard\_data <- read.csv("/Users/anshulkandpal/Downloads/be\_dataset/guardiandata.csv")

str(child\_data)

str(guard\_data)

view(child\_data)

view(guard\_data)

# we put all the datapoints which were only collected in baseline (eg, endline = 0) in this variable

baseline\_child\_data <- child\_data %>%

filter(endline == 0)

# mean and standard deviation of all the cognitive scores for treatment and control group

baseline\_child\_data %>%

group\_by(treatment) %>%

summarise(

N = n(),

across(c(age, qmat, bug, picture\_mem, vocab, anim\_code, receptive\_vocab, pict\_naming,

aser\_math, aser\_kannada, motor\_skills, body\_parts, colors, dial\_num, actions,

alpha, prob\_solve, copying, panamath, c\_totalscore\_all),

list(mean = ~mean(.x, na.rm = TRUE), sd = ~sd(.x, na.rm = TRUE)))

) %>% view()

# Define the variables for t-tests

test\_vars <- c("age", "qmat", "bug", "picture\_mem", "vocab", "anim\_code", "receptive\_vocab",

"pict\_naming", "aser\_math", "aser\_kannada", "motor\_skills", "body\_parts",

"colors", "dial\_num", "actions", "alpha", "prob\_solve", "copying", "panamath", "c\_totalscore\_all")

# Function to run t-test and extract key results

run\_t\_test <- function(var) {

t\_result <- t.test(baseline\_child\_data[[var]] ~ baseline\_child\_data$treatment, var.equal = TRUE)

data.frame(

Variable = var,

Mean\_Control = mean(baseline\_child\_data[[var]][baseline\_child\_data$treatment == 0], na.rm = TRUE),

Mean\_Treatment = mean(baseline\_child\_data[[var]][baseline\_child\_data$treatment == 1], na.rm = TRUE),

p\_value = t\_result$p.value

)

}

# Apply the function to all variables and store results in a dataframe

t\_test\_results <- map\_dfr(test\_vars, run\_t\_test)

# View results

print(t\_test\_results)

# OR

view(t\_test\_results)

# Now we tackle table pointer 2: "2. Please calculate the percentage of guardians-

# -who disagree about the type of primary school they are planning for their children"

guard\_data %>%

group\_by(prim\_type) %>%

summarise(

N = n()

)

# since we have 1776 '-99' values which are only used in endline, we clean this prim\_type data

# we basically don't need '-99' values for this analysis

clean\_guard\_data <- guard\_data %>%

filter(prim\_type %in% c("Private" , "Public"))

view(clean\_guard\_data)

disagreement\_data <- clean\_guard\_data %>%

group\_by(id\_child) %>%

summarise(

unique\_schools = n\_distinct(prim\_type)

) #in this command box, you basically count the unique preferences for schools for each child by their guardian, and increase count by 1 for each uniqueness..

# .. therefore, if there is uniqueness of 2, it means both guardians had a disagreement. Similary uniqueness of 1 means agreement

view(disagreement\_data)

n\_disagreements <- sum(disagreement\_data$unique\_schools == 2) # add up whenever there is a disagreement (unique schools = 2)

print(n\_disagreements) #print the count of disagreements amongst guardians

total\_children <- n\_distinct(clean\_guard\_data$id\_child)

print(total\_children)

percentage\_disagree <- (n\_disagreements / total\_children) \* 100 # using the formula of (disagreements/total) \* 100

print(percentage\_disagree)

# Now we tackle Pointer 3: "Please graph the distributions of the total score variable at baseline and endline by treatment status and interpret the graph"

# Scatterplot

child\_data %>%

ggplot(aes(x = c\_totalscore\_all, fill = factor(treatment))) +

geom\_density(alpha = 0.5) +

facet\_wrap(~ endline, labeller = labeller(endline = c("0" = "Baseline", "1" = "Endline"))) +

labs(title = "Total Score Distribution by Treatment Status",

x = "Total Score",

y = "Density",

fill = "Group") +

scale\_fill\_manual(values = c("0" = "blue", "1" = "red"),

labels = c("0" = "Control", "1" = "Treatment")) +

theme\_minimal()

# Boxplot

child\_data %>%

ggplot(aes(x = factor(treatment), y = c\_totalscore\_all, fill = factor(treatment))) +

geom\_boxplot(alpha = 0.5) +

facet\_wrap(~ endline, labeller = labeller(endline = c("0" = "Baseline", "1" = "Endline"))) +

labs(title = "Total Score Distribution by Treatment Status",

x = "Group",

y = "Total Score",

fill = "Group") +

scale\_fill\_manual(values = c("0" = "blue", "1" = "red"),

labels = c("0" = "Control", "1" = "Treatment")) +

theme\_minimal()

# OK, now we work on the 4th topic: "4. Please group the subtests thematically and create z-score...results in a few sentences."

# the themes have been made as following:

# 1. cognitive/non-verbal reasoning: qmat,bug search, picture memory, problem solving, panamath

# 2. language/verbal skills: vocabulary, receptive vocab, picture naming, alphabets, actions

# 3. quantitative(early academics): ASER math, DIAL number, ASER kannada

# 4. motor/visual motor/early concepts: animal coding, motor skills, copying, body parts, colors

# Compute z-scores for each theme and take the row-wise mean

child\_data$cognitive\_index <- rowMeans(scale(child\_data[, c("qmat", "bug", "picture\_mem", "prob\_solve", "panamath")]), na.rm = TRUE)

child\_data$language\_index <- rowMeans(scale(child\_data[, c("vocab", "receptive\_vocab", "pict\_naming", "alpha", "actions")]), na.rm = TRUE)

child\_data$quant\_index <- rowMeans(scale(child\_data[, c("aser\_math", "dial\_num", "aser\_kannada")]), na.rm = TRUE)

child\_data$motor\_index <- rowMeans(scale(child\_data[, c("anim\_code", "motor\_skills", "copying", "body\_parts", "colors")]), na.rm = TRUE)

# first, we filter the guardian\_data for females only

female\_guard\_data <- guard\_data %>%

filter(gender == "Female")

view(female\_guard\_data)

# now, we join the female guardian data and child data together, through left\_joining via child\_id to make a composite analysis table

analysis\_table <- merge(female\_guard\_data, child\_data)

# now we remove all the redundancies from analysis table

analysis\_table <- analysis\_table %>%

select(-qmat, -bug, -picture\_mem, -vocab, -anim\_code,

-receptive\_vocab, -pict\_naming, -aser\_math, -aser\_kannada,

-motor\_skills, -body\_parts, -colors, -dial\_num,

-actions, -alpha, -prob\_solve, -copying, -panamath, -hoursworked, -prim\_type)

view(analysis\_table)

# NOTE: FOR THIS FINAL ANALYSIS, I EXTRACTED THE DATA AND RE-ARRANGED IT IN EXCEL TO FACILITATE THE ANALYSIS

# I filtered baseline and endline cognitive scores accordingly and put them side-by-side

# The final\_analysis sheet is provided with the script

view(final\_analysis)

model\_cognitive\_index <- lm(cognitive\_index\_end ~ cognitive\_index\_base + age + edu + treatment, data = final\_analysis)

model\_language\_index <- lm(language\_index\_end ~ language\_index\_base + age + edu + treatment, data = final\_analysis)

model\_quantitative\_index <- lm(quant\_index\_end ~ quant\_index\_base + age + edu + treatment, data = final\_analysis)

model\_motor\_index <- lm(motor\_index\_end ~ motor\_index\_base + age + edu + treatment, data = final\_analysis)

summary(model\_cognitive\_index)

summary(model\_language\_index)

summary(model\_quantitative\_index)

summary(model\_motor\_index)