RCT Impact Evaluation

Victoria

2025-04-29

Evaluating the impact of an education intervention in Indonesia.

Reading data

```
install.packages("haven")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(haven)
students <- read_dta("RCT data/Data Files/students.dta")
```

Renaming and Lablelling Variables

```
install.packages("dplyr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
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
```

```
# Step 1: Rename variables
students <- students %>%
  rename(
    district_id        = var1,
    student_id         = var2,
    income_quintile    = var3,
    gender             = var4,
    dob_day            = var5,
    dob_month          = var6,
    dob_year           = var7,
    mother_edu         = var8,
    father_edu         = var9,
    literacy_baseline  = var10,
    literacy_endline   = var11,
    numeracy_score     = var12
  )

# Step 2: Label variables
attr(students$district_id,       "label") <- "District identifier"
attr(students$student_id,        "label") <- "Student identifier"
attr(students$income_quintile,   "label") <- "Income quintile"
attr(students$gender,            "label") <- "Student's gender"
attr(students$dob_day,           "label") <- "Student's date of birth - day"
attr(students$dob_month,         "label") <- "Student's date of birth - month"
attr(students$dob_year,          "label") <- "Student's date of birth - year"
attr(students$mother_edu,        "label") <- "Mother's education level"
attr(students$father_edu,        "label") <- "Father's education level"
attr(students$literacy_baseline, "label") <- "Literacy score, baseline"
attr(students$literacy_endline,  "label") <- "Literacy score, endline"
attr(students$numeracy_score,    "label") <- "Numeracy score"

# Assigning value labels
students$gender <- factor(students$gender,
                          levels = c(1, 5),
                          labels = c("Male", "Female"))

students$income_quintile <- factor(students$income_quintile,
                                   levels = 1:5,
                                   labels = c("Q1 (Lowest)", "Q2", "Q3", "Q4", "Q5 (Highest)"))

students$mother_edu <- factor(students$mother_edu,
                              levels = 1:4,
                              labels = c("No schooling", "Primary", "Secondary", "Tertiary and above"))

students$father_edu <- factor(students$father_edu,
                              levels = 1:4,
                              labels = c("No schooling", "Primary", "Secondary", "Tertiary and above"))
```

Summary Statistics

```
# Check summary statistics
install.packages("skimr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(skimr)
skim(students)
```

Data summary

Name	students
Number of rows	100000
Number of columns	12
________________________	
Column type frequency:	
character	2
factor	4
numeric	6

Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
district_id	0	1	4	4	0	73	0
student_id	0	1	9	9	0	100000	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
income_quintile	0	1	FALSE	5	Q1 : 20216, Q4: 20105, Q3: 19946, Q2: 19918
gender	0	1	FALSE	2	Mal: 52227, Fem: 47773
mother_edu	0	1	FALSE	4	Sec: 36892, No : 27530, Pri: 24465, Ter: 11113
father_edu	0	1	FALSE	4	Ter: 28067, Pri: 27545, Sec: 25210, No : 19178

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
dob_day	0	1	14.79	14.43	-99.00	8.00	16.00	24.00	99.00	▁▁▇▅▁
dob_month	0	1	6.34	3.04	-9.00	5.00	6.00	8.00	12.00	▁▁▂▇▆
dob_year	0	1	2010.00	1.16	2008.00	2009.00	2010.00	2011.00	2012.00	▃▅▇▅▂
literacy_baseline	0	1	21.30	2.53	-0.73	19.52	21.31	23.09	40.88	▁▁▇▁▁
literacy_endline	0	1	28.81	4.28	5.61	25.92	29.01	31.87	52.87	▁▂▇▁▁
numeracy_score	0	1	23.30	2.78	13.93	21.36	23.29	25.24	33.39	▁▅▇▃▁

Cleaning Data

```
# Checking for missing values
colSums(is.na(students))
```

```
##       district_id        student_id   income_quintile            gender
##                 0                 0                 0                 0
##           dob_day         dob_month          dob_year        mother_edu
##                 0                 0                 0                 0
##        father_edu literacy_baseline  literacy_endline    numeracy_score
##                 0                 0                 0                 0
```

```
# Check for negative dates
students %>% filter(dob_day <= 0 | dob_day > 31)
```

district_id <chr>	▸
1101	
1101	
1101	
1101	
1101	
1101	
1101	
1101	
1101	
1102	

1-10 of 1,007 rows | 1-1 of 12 columns   Previous **1** 2 Next

```
students %>% filter(dob_month <= 0 | dob_month > 12)
```

```
district_id                                                                  ▸
<chr>
1101
1101
1101
1101
1101
1101
1101
1101
1101
1101
1-10 of 1,036 rows | 1-1 of 12 columns                  Previous  1  2  Next
```

```
students %>% filter(dob_year < 1900 | dob_year > 2025)
```

```
0 rows | 1-1 of 12 columns
```

```
# Convert negative to absolute values
students <- students %>%
  mutate(
    dob_month = abs(dob_month),
  )
```

```
# Save dataset
write_dta(students,"students.dta")
```

Schools Dataset

```
library(haven)
schools <- read_dta("RCT data/Data Files/schools.dta")
```

Renaming and Labelling Variables

```
library(dplyr)

# Step 1: Rename the variables
schools <- schools %>%
  rename(
    district_id     = var1,
    school_id       = var2,
    treatment       = var3,
    num_teachers    = var4,
    num_classrooms  = var5,
    urban_rural     = var6
  )

# Step 2: Label each variable
attr(schools$district_id,    "label") <- "District identifier"
attr(schools$school_id,      "label") <- "School identifier"
attr(schools$treatment,      "label") <- "Treatment assignment"
attr(schools$num_teachers,   "label") <- "Number of teachers in the school"
attr(schools$num_classrooms, "label") <- "Number of classrooms in the school"
attr(schools$urban_rural,    "label") <- "Urban/rural status"

# Assigning value labels
schools $urban_rural <- factor(schools$urban_rural,
                               levels = c(1, 7),
                               labels = c("Urban", "Rural"))
schools$treatment <- factor(schools$treatment,
                             levels = 0:2,
                             labels = c("Control Group", "Treatment 1", "Treatment 2"))
```

Summary Statistics

```
library(skimr)
skim(schools)
```

Data summary

Name	schools
Number of rows	1156
Number of columns	6
_______________________	
Column type frequency:	
character	2
factor	2
numeric	2
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
district_id	0	1	4	4	0	73	0
school_id	0	1	6	6	0	1000	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
treatment	0	1	FALSE	3	Con: 386, Tre: 385, Tre: 385
urban_rural	0	1	FALSE	2	Rur: 831, Urb: 325

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
num_teachers	0	1.00	2.88	6.01	-9	2	3	5	38	▂▇▁▁▁
num_classrooms	156	0.87	-423.11	507.38	-999	-999	-92	-92	1302	▇▇▃▁▁

```
## Identifying extreme or negative values
min(schools$num_classrooms, na.rm = TRUE)
```

```
## [1] -999
```

```
max(schools$num_classrooms, na.rm = TRUE)
```

```
## [1] 1302
```

```
min(schools$num_teachers, na.rm = TRUE)
```

```
## [1] -9
```

```
max(schools$num_teachers, na.rm = TRUE)
```

```
## [1] 38
```

Cleaning Data

We seem to have negative teachers and classrooms, we’ll convert these to absolute values

```
schools <- schools %>%
  mutate(
    num_teachers   = abs(num_teachers),
    num_classrooms = abs(num_classrooms)
  )
write_dta(schools, "schools.dta")
```

Importing additional CSV data file

```
library(readr)
take_up <- read_csv("RCT data/Data Files/take-up.csv")
```

```
## Rows: 100000 Columns: 3
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): school
## dbl (2): takeup, student
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Skimming the takeup data

```
library(skimr)
skim(take_up)
```

Data summary

Name	take_up
Number of rows	100000
Number of columns	3
_______________________	
Column type frequency:	
character	1
numeric	2
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
school	0	1	6	10	0	2453	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
takeup	0	1	0.96	0.82	0	0.00	1	2	2	▇▁▇▁▇
student	0	1	73.07	65.55	1	25.75	52	102	422	▇▂▁▁▁

Cleaning the Take up dataset

```
# Remove trailing spaces, extra zeros and ashes in the school variable
take_up$school <- trimws(take_up$school)
take_up$school <- gsub("^0+", "", take_up$school)
take_up$school <- gsub("-", "", take_up$school)
take_up$school <- gsub(" ", "", take_up$school)
```

Merging Datasets

```
# Aggregating schools data to avoid many to many merging
agg_schools <- schools %>%
  group_by(district_id) %>%
  summarize(
    avg_teachers = mean(num_teachers, na.rm = TRUE),   # Average number of teachers per district
    avg_classrooms = mean(num_classrooms, na.rm = TRUE), # Average number of classrooms
    urban_ratio = mean(urban_rural == "Urban", na.rm = TRUE), # Proportion of urban schools
    total_schools = n(), # Total number of schools in the district
    treatment_assignment = first(treatment) # Keep treatment info
  )

# Merge agg schools and students dataset
students_schools <- merge(students, agg_schools, by = "district_id", all.x = TRUE)
write_dta(students_schools, "students_schools.dta")

library(skimr)
skim(students_schools)
```

Data summary

Name	students_schools
Number of rows	100000
Number of columns	17
_______________________	
Column type frequency:	
character	2
factor	5
numeric	10
________________________	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
district_id	0	1	4	4	0	73	0
student_id	0	1	9	9	0	100000	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
income_quintile	0	1	FALSE	5	Q1 : 20216, Q4: 20105, Q3: 19946, Q2: 19918
gender	0	1	FALSE	2	Mal: 52227, Fem: 47773
mother_edu	0	1	FALSE	4	Sec: 36892, No : 27530, Pri: 24465, Ter: 11113
father_edu	0	1	FALSE	4	Ter: 28067, Pri: 27545, Sec: 25210, No : 19178
treatment_assignment	0	1	FALSE	3	Tre: 37367, Tre: 32047, Con: 30586

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
dob_day	0	1	14.79	14.43	-99.00	8.00	16.00	24.00	99.00	▁▁▇▅▁
dob_month	0	1	6.52	2.62	1.00	5.00	7.00	8.00	12.00	▃▆▇▆▃
dob_year	0	1	2010.00	1.16	2008.00	2009.00	2010.00	2011.00	2012.00	▃▅▇▅▂
literacy_baseline	0	1	21.30	2.53	-0.73	19.52	21.31	23.09	40.88	▁▁▇▁▁
literacy_endline	0	1	28.81	4.28	5.61	25.92	29.01	31.87	52.87	▁▂▇▁▁
numeracy_score	0	1	23.30	2.78	13.93	21.36	23.29	25.24	33.39	▁▅▇▃▁
avg_teachers	0	1	5.44	1.06	3.75	4.60	5.24	5.93	8.67	▆▇▅▂▁
avg_classrooms	0	1	463.57	126.76	209.00	357.77	463.36	550.79	738.43	▃▇▆▆▂
urban_ratio	0	1	0.27	0.45	0.00	0.00	0.00	1.00	1.00	▇▁▁▁▃

total_schools	0	1	15.86	1.36	13.00	15.00	16.00	17.00	19.00	▂▇▂▅▂

Mapping

```
# Installing required packages
install.packages( "ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(dplyr)
library(ggplot2)
install.packages("sf")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(sf)
```

```
## Linking to GEOS 3.8.0, GDAL 3.0.4, PROJ 6.3.1; sf_use_s2() is TRUE
```

```
install.packages("tmap")
```

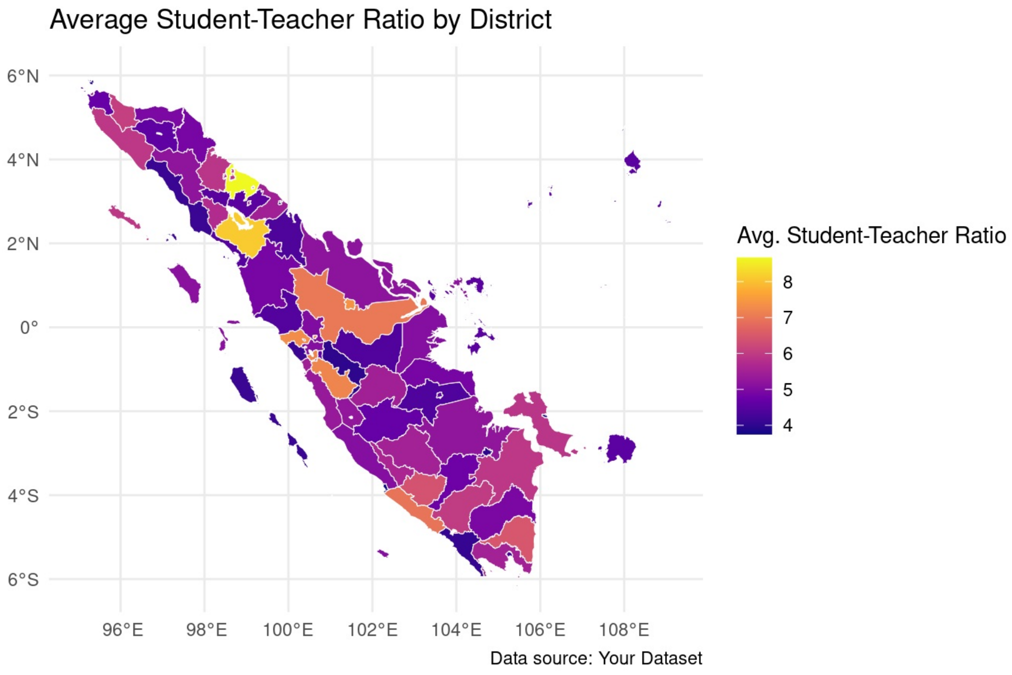
```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
## Warning in install.packages("tmap"): installation of package 'tmap' had
## non-zero exit status
```

```
# reading map data
district_map <- st_read("RCT data/Data Files/sumatra.shp")
```

```
## Reading layer `sumatra' from data source
##   `/cloud/project/RCT data/Data Files/sumatra.shp' using driver `ESRI Shapefile'
## Simple feature collection with 73 features and 2 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 95.00708 ymin: -6.172917 xmax: 109.1663 ymax: 6.08125
## Geodetic CRS:  WGS 84
```

```
# Calculating student teacher ratio
library(dplyr)

district_summary <- students_schools %>%
  group_by(district_id) %>%
  summarise(
    num_students = n_distinct(student_id),
    avg_teachers = first(avg_teachers)
  ) %>%
  mutate(student_teacher_ratio = num_students / avg_teachers)

# Merge with map data

library(dplyr)

district_map <- district_map %>%
  rename(district_id = KAB)
map_data <- district_map %>%
  left_join(district_summary, by = "district_id")

#Plotting on Map
library(ggplot2)
library(sf)


ggplot(data = map_data) +
  geom_sf(aes(fill = avg_teachers), color = "white") +
  scale_fill_viridis_c(option = "C", name = "Avg. Student-Teacher Ratio") +
  labs(
    title = "Average Student-Teacher Ratio by District",
    caption = "Data source: Your Dataset"
  ) +
  theme_minimal()
```


Average Student-Teacher Ratio by District
6°N
4°N
2°N
0°
2°S
4°S
6°S
96°E
98°E
100°E
102°E
104°E
106°E
108°E
Avg. Student-Teacher Ratio
8
7
6
5
4
Data source: Your Dataset

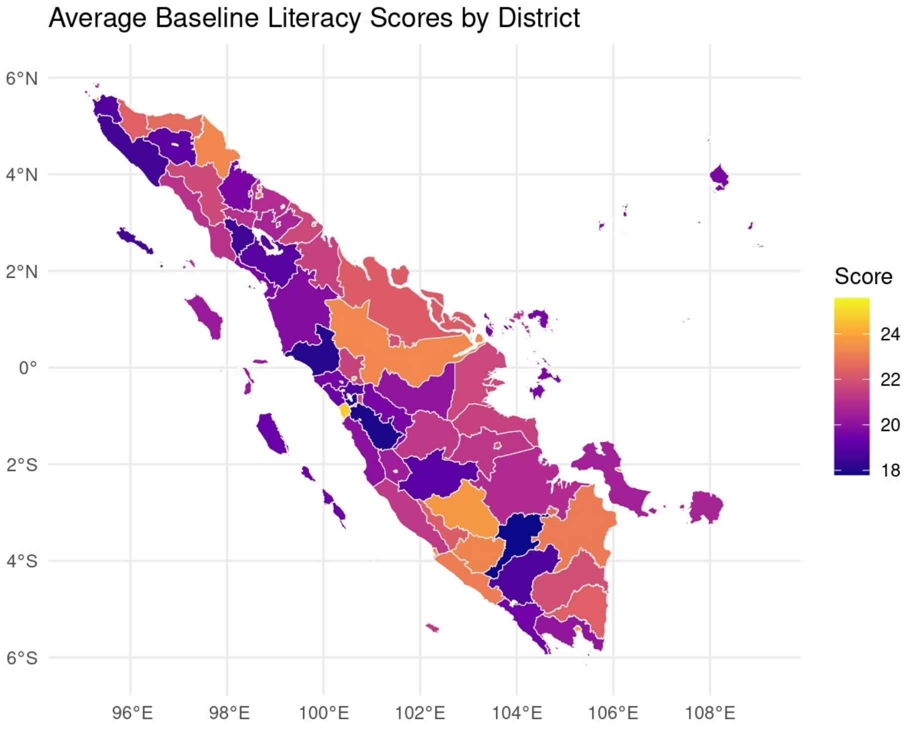
### Literacy Scores Before and After

Treatment

```
# Loading required libraries
library(dplyr)
library(ggplot2)
library(sf)

# 1. Reading in the shapefile
district_map <- st_read("RCT data/Data Files/sumatra.shp")
```

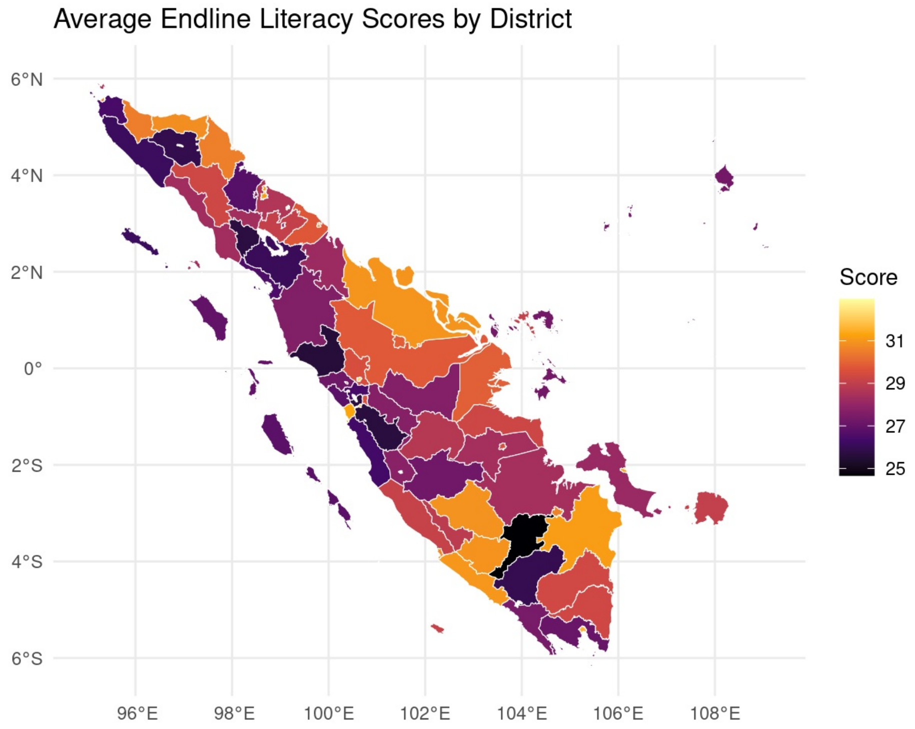
```
## Reading layer `sumatra' from data source
##   `/cloud/project/RCT data/Data Files/sumatra.shp' using driver `ESRI Shapefile'
## Simple feature collection with 73 features and 2 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 95.00708 ymin: -6.172917 xmax: 109.1663 ymax: 6.08125
## Geodetic CRS:  WGS 84
```

```
# 2. Summarizing average scores by district
literacy_summary <- students_schools %>%
  group_by(district_id) %>%
  summarise(
    avg_lit_base = mean(literacy_baseline, na.rm = TRUE),
    avg_lit_end = mean(literacy_endline, na.rm = TRUE)
  )

#Renaming KAB to district_id
district_map <- district_map %>%
  rename(district_id = KAB)

# 4. Merging literacy summary with spatial data
map_data <- left_join(district_map, literacy_summary, by = "district_id")

# 5. Plotting baseline literacy map
ggplot(map_data) +
  geom_sf(aes(fill = avg_lit_base), color = "white") +
  scale_fill_viridis_c(option = "plasma", na.value = "grey90") +
  theme_minimal() +
  labs(
    title = "Average Baseline Literacy Scores by District",
    fill = "Score"
  )
```



Average Baseline Literacy Scores by District

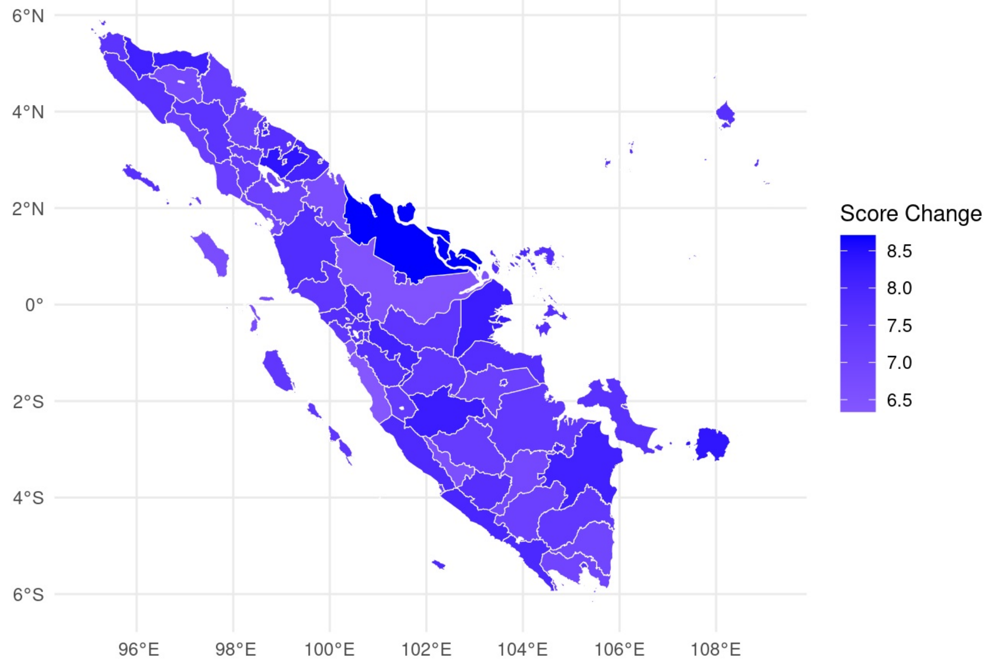
```
# 6. Plotting endline literacy map
ggplot(map_data) +
  geom_sf(aes(fill = avg_lit_end), color = "white") +
  scale_fill_viridis_c(option = "inferno", na.value = "grey90") +
  theme_minimal() +
  labs(
    title = "Average Endline Literacy Scores by District",
    fill = "Score"
  )
```



Change in literacy scores

```
# 1. Calculating average change per district
literacy_diff <- students_schools %>%
  group_by(district_id) %>%
  summarise(
    avg_score_change = mean(literacy_endline - literacy_baseline, na.rm = TRUE)
  )

# 2. Merging with shapefile data
map_diff <- left_join(district_map, literacy_diff, by = "district_id")

# 3. Plotting the map
ggplot(map_diff) +
  geom_sf(aes(fill = avg_score_change), color = "white") +
  scale_fill_gradient2(
    low = "red",
    mid = "white",
    high = "blue",
    midpoint = 0,
    na.value = "grey90"
  ) +
  theme_minimal() +
  labs(
    title = "Change in Literacy Scores (Endline - Baseline) by District",
    fill = "Score Change"
  )
```

Change in Literacy Scores (Endline - Baseline) by District


6°N
4°N
2°N
0°
2°S
4°S
6°S
96°E
98°E
100°E
102°E
104°E
106°E
108°E
Score Change
8.5
8.0
7.5
7.0
6.5