# Quatlity Eduction and the Connection to Gender Equality in Africa

Ada Fu, Yuting Jiang, Ada Wang, and Sark Asadourian

## Question 1:

**Research question: Are there specific groups of African countries with similar Quality Education SDG scores and Gender Inequality Index(GII) scores?** Main methods: K-means clustering

1. read all files

```
country_code <- read_csv("country_codes.csv")

indicator <- read_csv("country_indicators.csv")

SDG_data <- read_csv("sdr_fd5e4b5a.csv")
```

2. data wrangling

```
indicators <- indicator %>%
  select(-...1) %>%  # remove first column
  select(iso3, everything()) %>%  # reorder the columns to put iso3 as column 1
  rename(country_code_iso3 = iso3)  # rename first column to country_code_iso3

names(indicators)[names(indicators)=="hdr_gii_2021"] <-
  "gii_2021" # rename the column to gii_2021

c_indicator <- indicators %>%
  select(country_code_iso3, gii_2021) # select the columns we need
```

```
names(SDG_data)[names(SDG_data)=="Goal 4 Score"] <-
  "Goal_4_score" # rename the column to Goal_4_score

sdg <- SDG_data %>%
  select(-...1) %>% # remove first column
  select(country_label, Goal_4_score) # select the columns we need
```

```
clean_sdg <- sdg %>% # clean missing sdg data
  filter(!is.na(country_label)) %>%
  filter(!is.na(Goal_4_score))

clean_indicator <- c_indicator %>% # cleaning missing indicator data
  filter(!is.na(country_code_iso3)) %>%
  filter(!is.na(gii_2021))
```

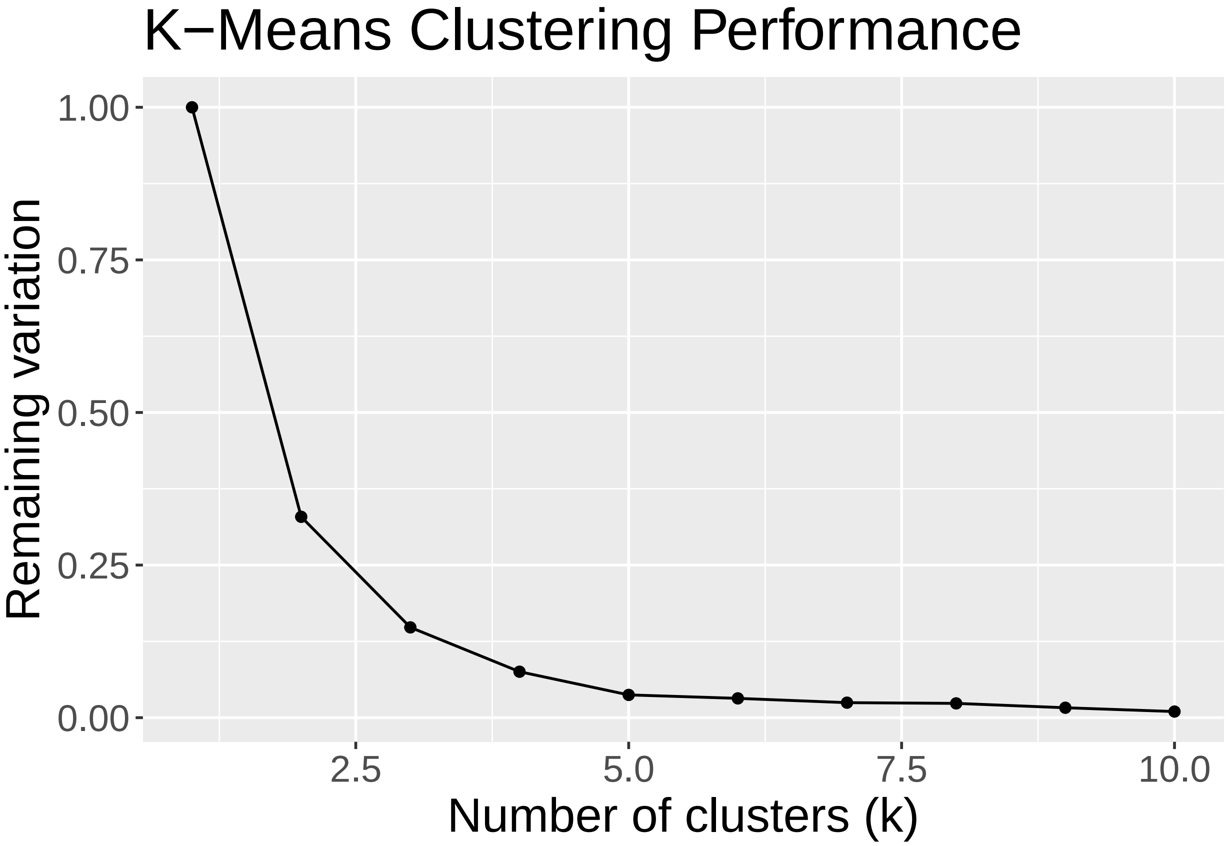
```
# rename the columns
names(country_code)[names(country_code)=="Country or Area_en (M49)"] <-
  "country_or_area"
names(country_code)[names(country_code)=="Region Name_en (M49)"] <-
  "region_name"
names(country_code)[names(country_code)=="ISO-alpha3 Code (M49)"] <-
```

```
  "iso3"

code <- country_code %>%
  select(-...1) %>% # remove first column
  select(region_name, iso3, country_or_area) # select the columns we need

clean_code <- code %>%
  filter(region_name == "Africa") # filter Africa countries
```

```
# combine tables
data1 <- right_join(x=clean_indicator, y=clean_code, by=c("country_code_iso3"=
                                                          "iso3"))
data2 <- right_join(x=clean_sdg, y=clean_code, by=c("country_label"=
                                                          "country_or_area"))
final_data <- inner_join(x=data1, y=data2,
                         by=c("country_code_iso3"="iso3",
                                "region_name"="region_name",
                                "country_or_area"="country_label"))

f_data <- final_data %>% # reorder the columns
  select(region_name, country_code_iso3, country_or_area, gii_2021, Goal_4_score)

ff_data <- f_data %>% # cleaning missing data
  filter(!is.na(gii_2021)) %>%
  filter(!is.na(Goal_4_score))
```

3. perform the kmeans clustering

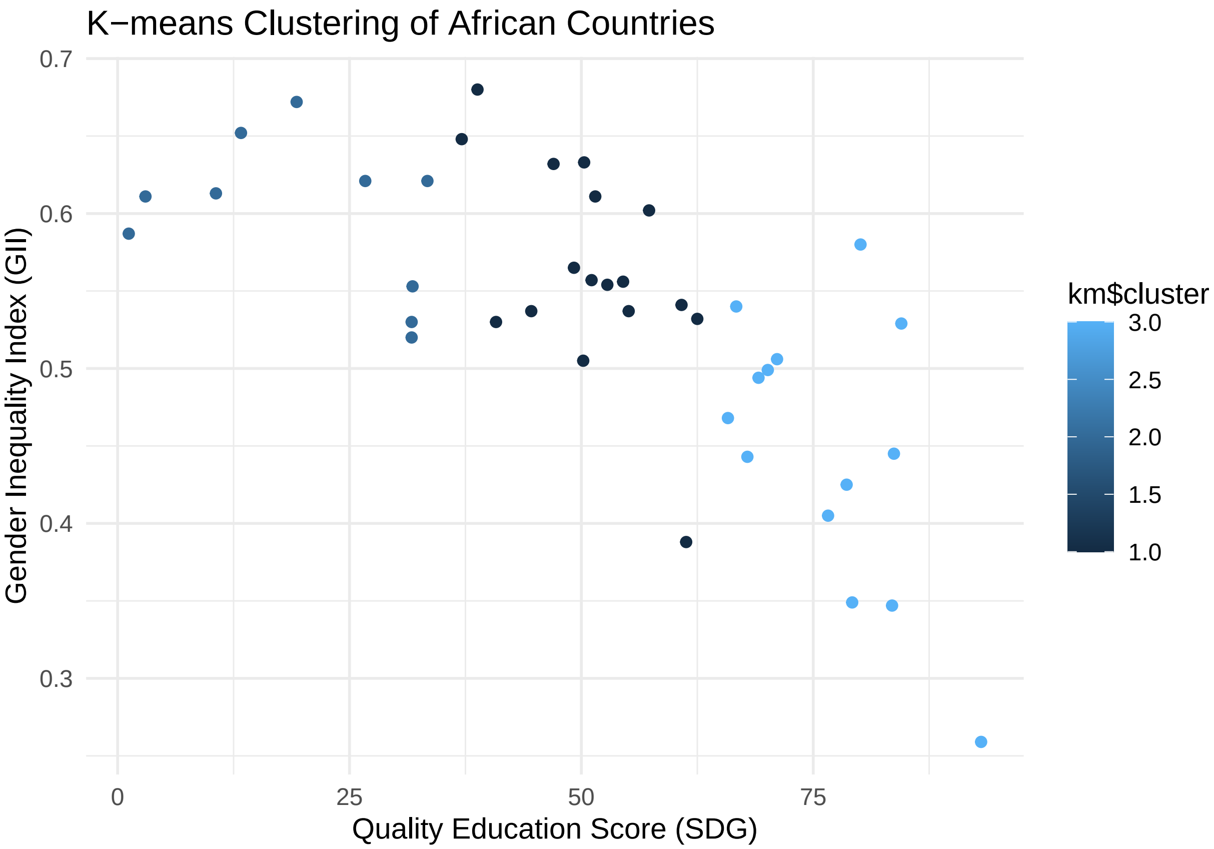
```
cluster_data <- ff_data %>%
  select(gii_2021, Goal_4_score) # select the columns we need for ploting

explained_ss <- rep(NA, 10)
# Perform K-means clustering for different values of k
for (k in 1:10) {
  clustering <- kmeans(cluster_data, centers = k)
  explained_ss[k] <- clustering$betweenss / clustering$totss
}
# Plot the Elbow Method
ggplot() +
aes(x=1:10, y=1-explained_ss) +
  geom_line() +
  geom_point() +
  labs(x="Number of clusters (k)",
       y="Remaining variation",
       title="K-Means Clustering Performance") +
  theme(text=element_text(size=18))
```



```r
set.seed(130) # for reproducibility
k <- 3 # choose the number of clusters
km <- kmeans(cluster_data, k)
```

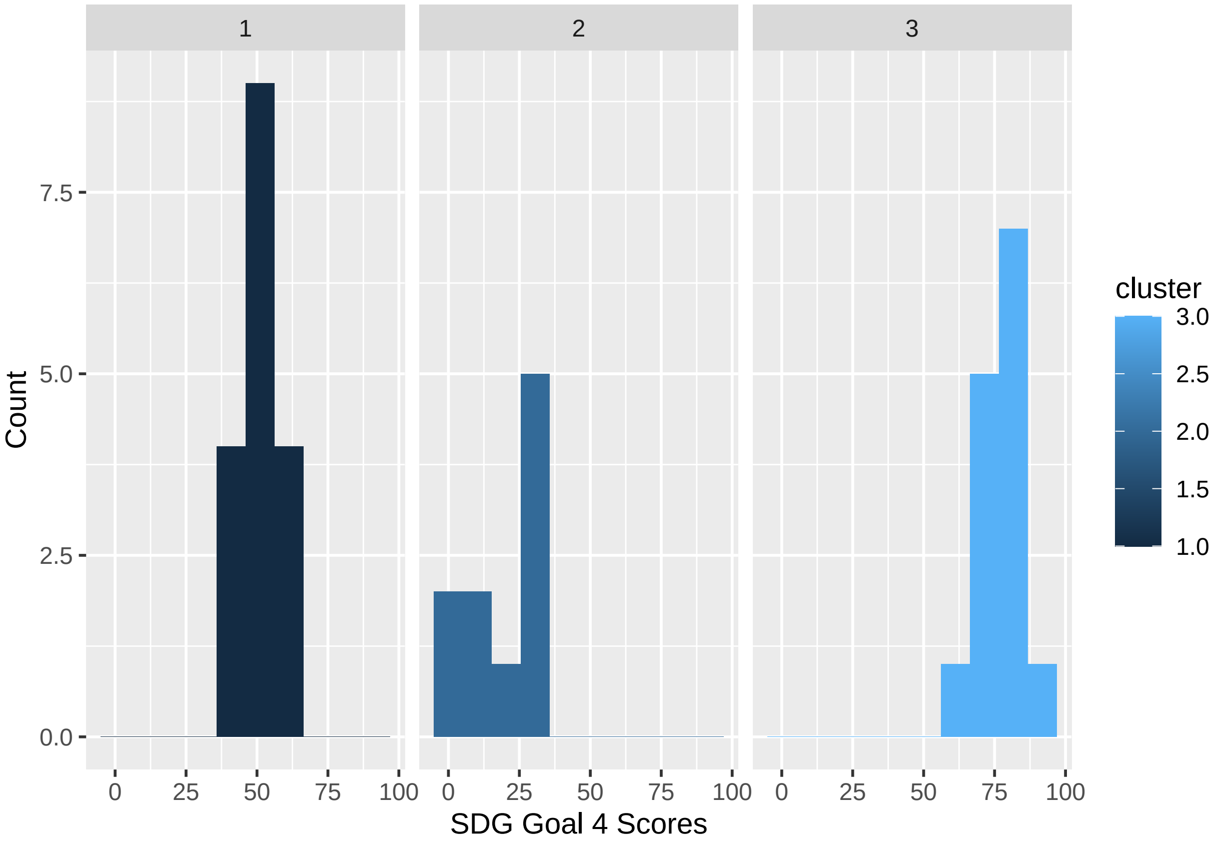
4. Visualize the clusters using a scatter plot

```r
ggplot(cluster_data, aes(x = Goal_4_score, y = gii_2021, color = km$cluster)) + geom_point() +
  labs(title = "K-means Clustering of African Countries",
       x = "Quality Education Score (SDG)",
       y = "Gender Inequality Index (GII)") +
  theme_minimal()
```



5. Create histograms to compare the SDG4 scores across clusters

```
cluster_data <- cluster_data %>%
  mutate(cluster = km$cluster)

# create a histogram for SDG4 score score
cluster_data %>% ggplot(aes(x=Goal_4_score, group=cluster, fill=cluster)) +
  geom_histogram(bins=10) +
  labs(x="SDG Goal 4 Scores",
       y="Count") +
  facet_wrap(~cluster)
```



```
  theme(text=element_text(size=18))
```

```
## List of 1
##  $ text:List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : num 18
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : NULL
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

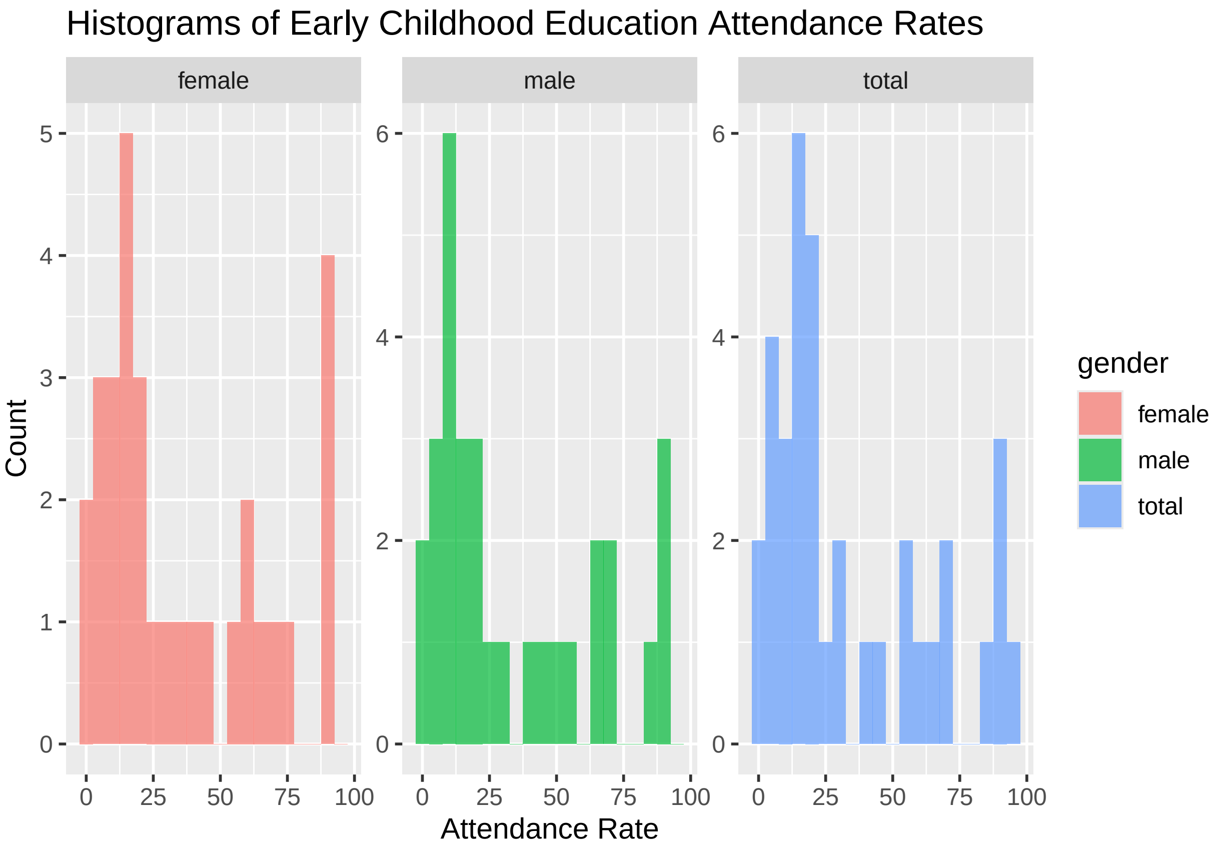
```
avg_scores <- cluster_data %>% # make a summary table to see the average scores
  group_by(cluster) %>%
  summarise(avg_gii_rank = mean(gii_2021),
            avg_Goal_4_score = mean(Goal_4_score))
```

## Question 2:

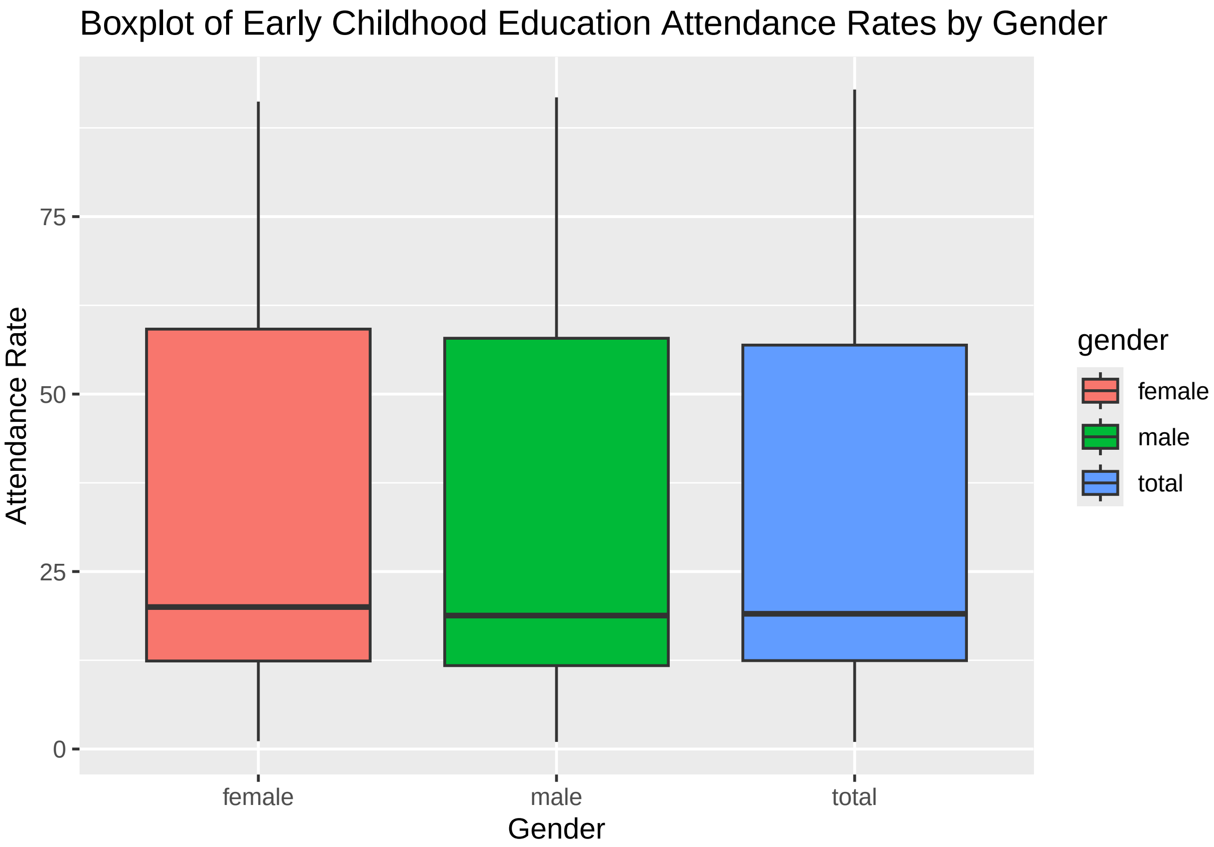
**Research question: In African countries, are there differents in early childhood education attendance rates across males and females from 2013 to 2021?** Main methods: Hypothesis Testing

```
#Read the country codes and indicators data
country_indicators <- read_csv("country_indicators.csv")
country_codes <- read_csv("country_codes.csv")

#Rename 'Region Code (M49)' to 'Region_name'
country_codes <- rename(country_codes, Region_name = "Region Code (M49)")

#Filter for the region code '2' which we assume represents Africa
african_region_codes <- filter(country_codes, Region_name == 2)

#Merge two data with the same region name as Africa
african_region_indicators <- inner_join(country_indicators, african_region_codes)

#Select columns of interest and transform data for plotting
columns_of_interest <- c(
  'sowc_early-childhood-development__attendance-in-early-childhood-education-2013-2021-r_total',
  'sowc_early-childhood-development__attendance-in-early-childhood-education-2013-2021-r_male',
  'sowc_early-childhood-development__attendance-in-early-childhood-education-2013-2021-r_female'
)
```

```
#Extract the necessary data for the African region
ecd_attendance_data <- african_region_indicators[columns_of_interest]

#Transform the data for visualization
ecd_attendance_data_long <- pivot_longer(
  ecd_attendance_data,
  cols = starts_with('sowc_early-childhood-development__attendance-in-early-childhood-education-2013-20
  names_to = 'gender',
  values_to = 'attendance_rate'
)

#Modify the 'gender' column to have cleaner names
ecd_attendance_data_long$gender <- sub('sowc_early-childhood-development__attendance-in-early-childhood-

#Plot histograms and boxplots
ggplot(data = ecd_attendance_data_long, aes(x = attendance_rate, fill = gender)) +
  geom_histogram(binwidth = 5, alpha = 0.7) +
  facet_wrap(~gender, scales = 'free_y') +
  labs(title = 'Histograms of Early Childhood Education Attendance Rates', x = 'Attendance Rate', y = '(
```



```
ggplot(data = ecd_attendance_data_long, aes(x = gender, y = attendance_rate, fill = gender)) +
  geom_boxplot() +
  labs(title = 'Boxplot of Early Childhood Education Attendance Rates by Gender', x = 'Gender', y = 'At
```



Hypothesis testing:

```
#Filter the data for only male and female
filtered_data <- ecd_attendance_data_long %>%
  filter(gender %in% c("male", "female"))

# Perform a two-sample t-test
t_test_results <- t.test(
  attendance_rate ~ gender,
  data = filtered_data,
  alternative = "two.sided", # to test for any difference in means
  mu = 0,                    # the difference in means under the null hypothesis
  paired = FALSE,            # set to FALSE because the samples are independent
  var.equal = FALSE          # set to FALSE to perform Welch's t-test
)

# Output the results
print(t_test_results)
```

```
##
##  Welch Two Sample t-test
##
## data:  attendance_rate by gender
## t = 0.19712, df = 59.999, p-value = 0.8444
## alternative hypothesis: true difference in means between group female and group male is not equal to
## 95 percent confidence interval:
##  -13.82817  16.85152
## sample estimates:
```

```
## mean in group female   mean in group male
##             35.72587             34.21419
```

## Question 3:

**Research question: Among African countries, what's the relationship between quality education SDG progress and gender equality SDG progress?** Main methods: linear regression

1. read all files

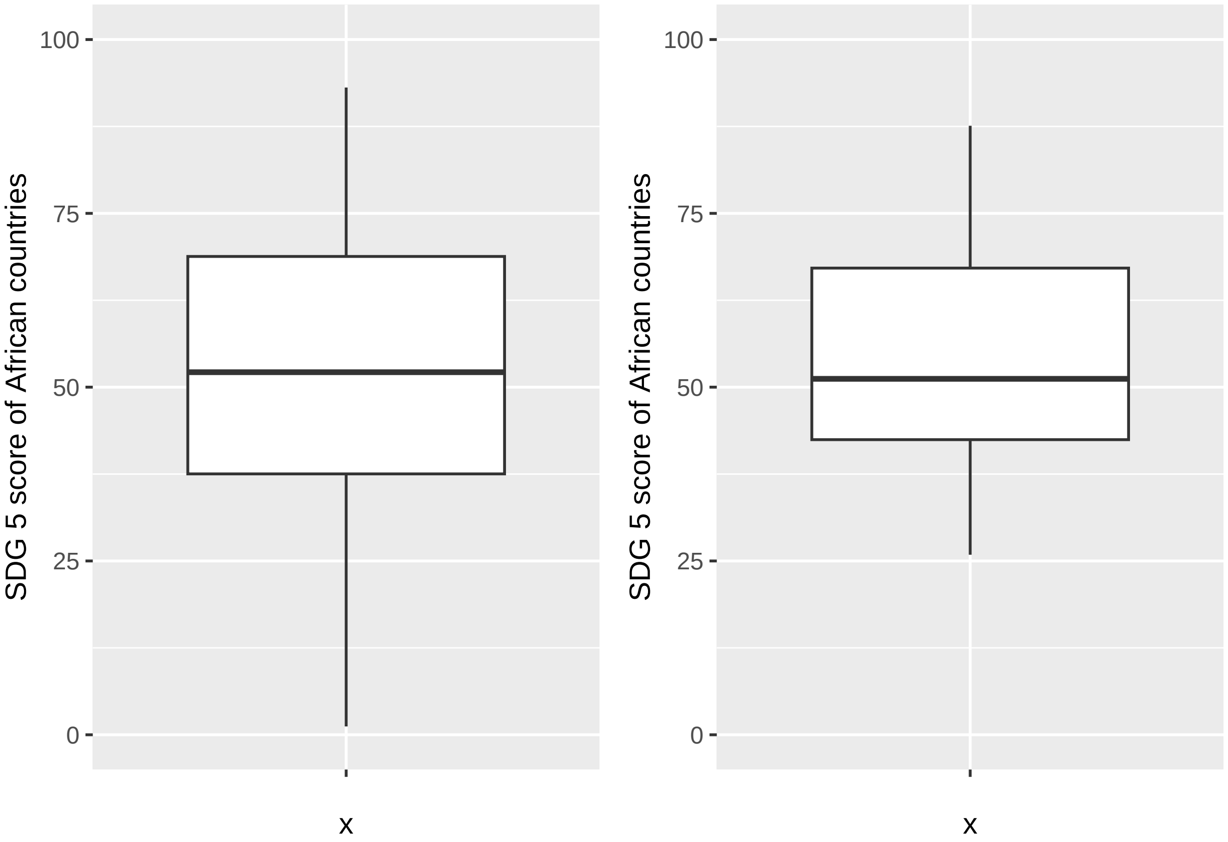
```
country_name <- read_csv("country_codes.csv")

SDG_score <- read_csv("sdr_fd5e4b5a.csv")
```

2. data wrangling

```
#select and filter out all the African countries in country_name
#rename
names(country_name)[names(country_name)=="Country or Area_en (M49)"] <-
  "country_area_name"
names(country_name)[names(country_name)=="Region Name_en (M49)"] <-
  "region_name"
cleaned_name <- country_name %>% select(country_area_name,
                                        region_name) %>%
  filter(region_name == "Africa")

names(SDG_score)[names(SDG_score)=="Goal 4 Score"] <-
  "Goal_4_score"
names(SDG_score)[names(SDG_score)=="Goal 5 Score"] <-
  "Goal_5_score"

SDG4_5 <- SDG_score %>% select(Goal_4_score, Goal_5_score, country_label)

#integrate the two data set by inner join
integrated_data <- right_join(x=SDG4_5, y=cleaned_name, by= c("country_label"= "country_area_name")) %>%
```

3. see distributions in both SDG 4 score and SDG 5 score

```
plot1 <- integrated_data %>% ggplot(aes(x="", y=Goal_4_score))+
  geom_boxplot()+
  labs(y="SDG 5 score of African countries")

plot2 <- integrated_data %>% ggplot(aes(x= "", y=Goal_5_score))+
  geom_boxplot()+
  labs(y="SDG 5 score of African countries")

plot1 <- plot1 + scale_y_continuous(limits = c(0, 100))

plot2 <- plot2 + scale_y_continuous(limits = c(0, 100))

grid.arrange(plot1, plot2, ncol=2)
```

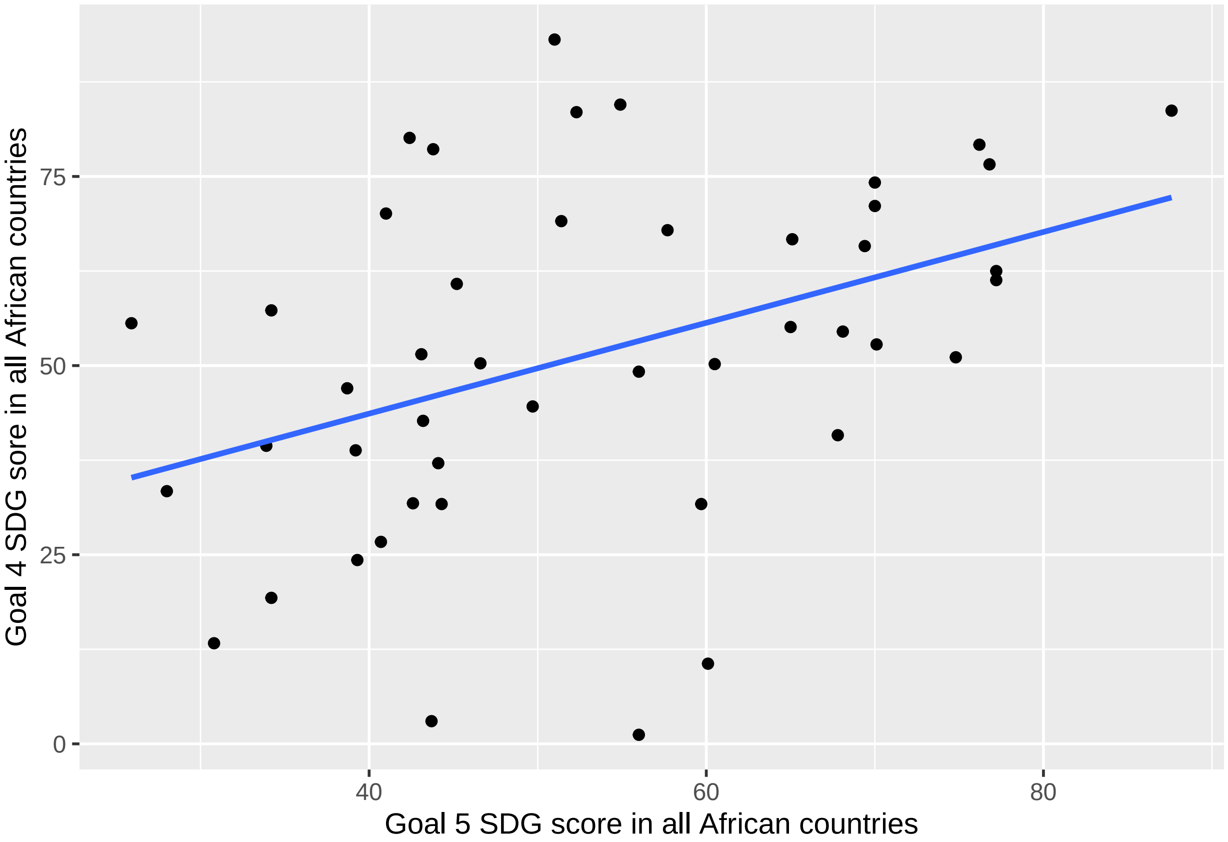


4. linear regression model

```
#regression coefficient
model <- lm(Goal_4_score ~ Goal_5_score, data = integrated_data)
summary(model) $ coefficients
```

```
##                Estimate Std. Error  t value   Pr(>|t|)
## (Intercept)  19.6342689 11.2267728 1.748879 0.08728561
## Goal_5_score  0.6003771  0.2026965 2.961951 0.00491416
```

```
#scatterplot
integrated_data %>% ggplot(aes(x=Goal_5_score, y=Goal_4_score)) + geom_point()+
  labs(x="Goal 5 SDG score in all African countries",
       y= "Goal 4 SDG sore in all African countries")+
  geom_smooth(method= "lm", se = FALSE)
```



**calculating the r value

```
##Since we have Na value in observations, so we first filter the data
data2 <- integrated_data %>% filter(!is.na(Goal_4_score) & !is.na(Goal_5_score))
##r value
cor(x= data2$Goal_5_score, y= data2$Goal_4_score)
```

```
## [1] 0.4077287
```