STA130 Final Project – Winter 2024

Om Patel, Kabir Kumar, Tianqi Zhang, Jin Masuda, Yianni Culmone

Guiding Question

How well are South American countries progressing toward achieving their SDG goals?

Research Question 1

Can we observe any trends in SDG velocity rates by categorizing countries based on their education levels?

Step 1

Loading data

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

# List of specified countries
countries <- c("Brazil", "Colombia", "Argentina", "Peru", "Venezuela", "Chile",
               "Ecuador", "Bolivia", "Paraguay", "Uruguay", "Guyana", "Suriname")

# Filter the dataset for the specified countries and select the necessary columns
country_codes <- country_codes %>%
  filter(`Country or Area_en (M49)` %in% countries) %>%
  select(`Country or Area_en (M49)`, `ISO-alpha3 Code (M49)`)

# View the filtered data
print(country_codes)
```

```
## # A tibble: 10 x 2
##    ‘Country or Area_en (M49)‘ ‘ISO-alpha3 Code (M49)‘
##    <chr>                      <chr>
##  1 Argentina                  ARG
##  2 Brazil                     BRA
##  3 Chile                      CHL
##  4 Colombia                   COL
##  5 Ecuador                    ECU
##  6 Guyana                     GUY
##  7 Paraguay                   PRY
##  8 Peru                       PER
##  9 Suriname                   SUR
## 10 Uruguay                    URY
```

Step 2

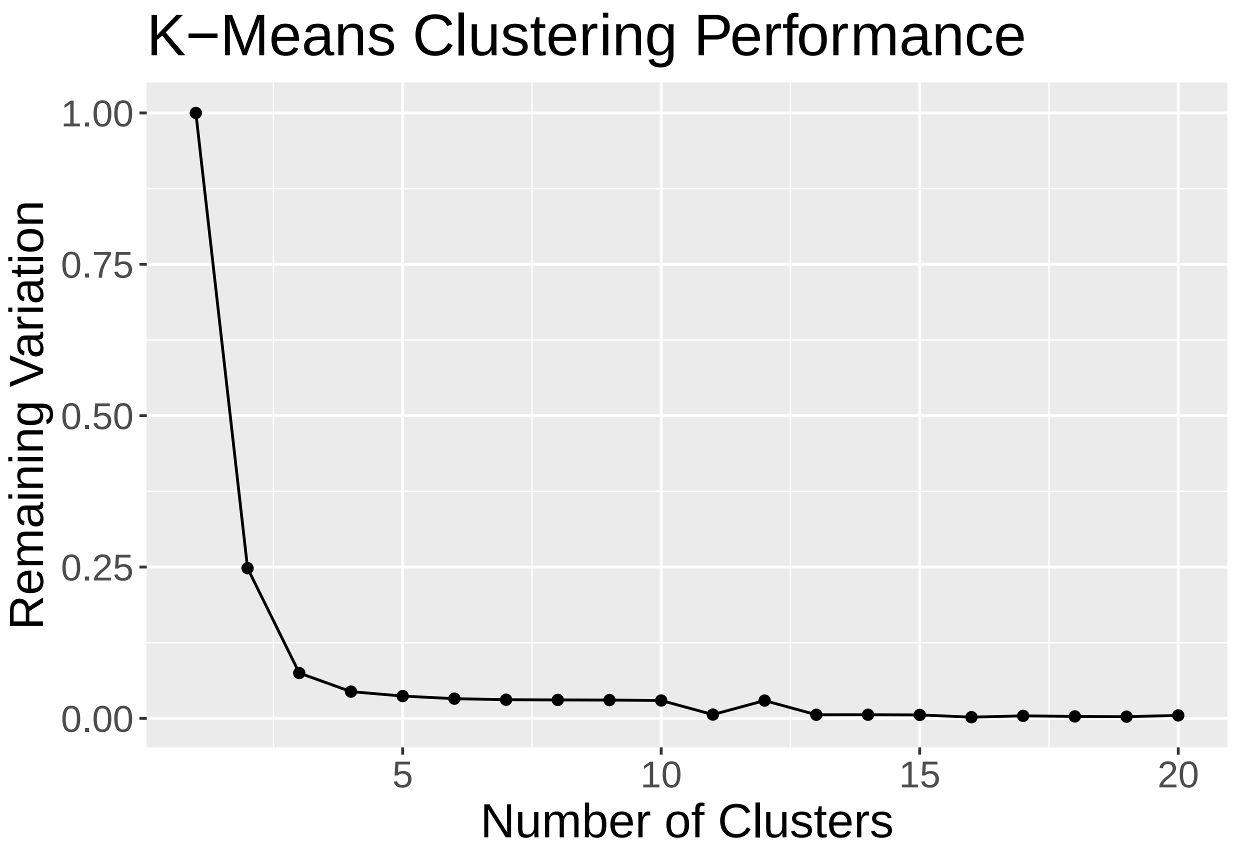
Filter relevant columns and handle missing data

```
country_indicators <-
  read_csv("country_indicators.csv") %>%
  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

filtered_data <- country_indicators %>%
  select(country_code_iso3, contains("sowc_education__completion_completion-rate-2013-2021-r_primary-edu

filtered_data <- filtered_data %>% filter(!is.na(filtered_data$`sowc_education__completion_completion-ra
```

Step 3

Optimal number of clusters with the elbow method

```
# Step 3: Optimal number of clusters with the elbow method
set.seed(100)
explained_ss <- rep(NA, 20)
for(k in 1:20){
  # run k-means on the data
  clustering <- kmeans(filtered_data$`sowc_education__completion_completion-rate-2013-2021-r_primary-edu
  explained_ss[k] <- clustering$betweenss / clustering$totss
}

# Plot evolution of metric as a function of k
ggplot() +
  aes(x=1:20, y=1-explained_ss) +
  geom_line() +
  geom_point() +
  labs(x="Number of Clusters",
       y="Remaining Variation",
       title="K-Means Clustering Performance") +
  theme(text=element_text(size=18))
```

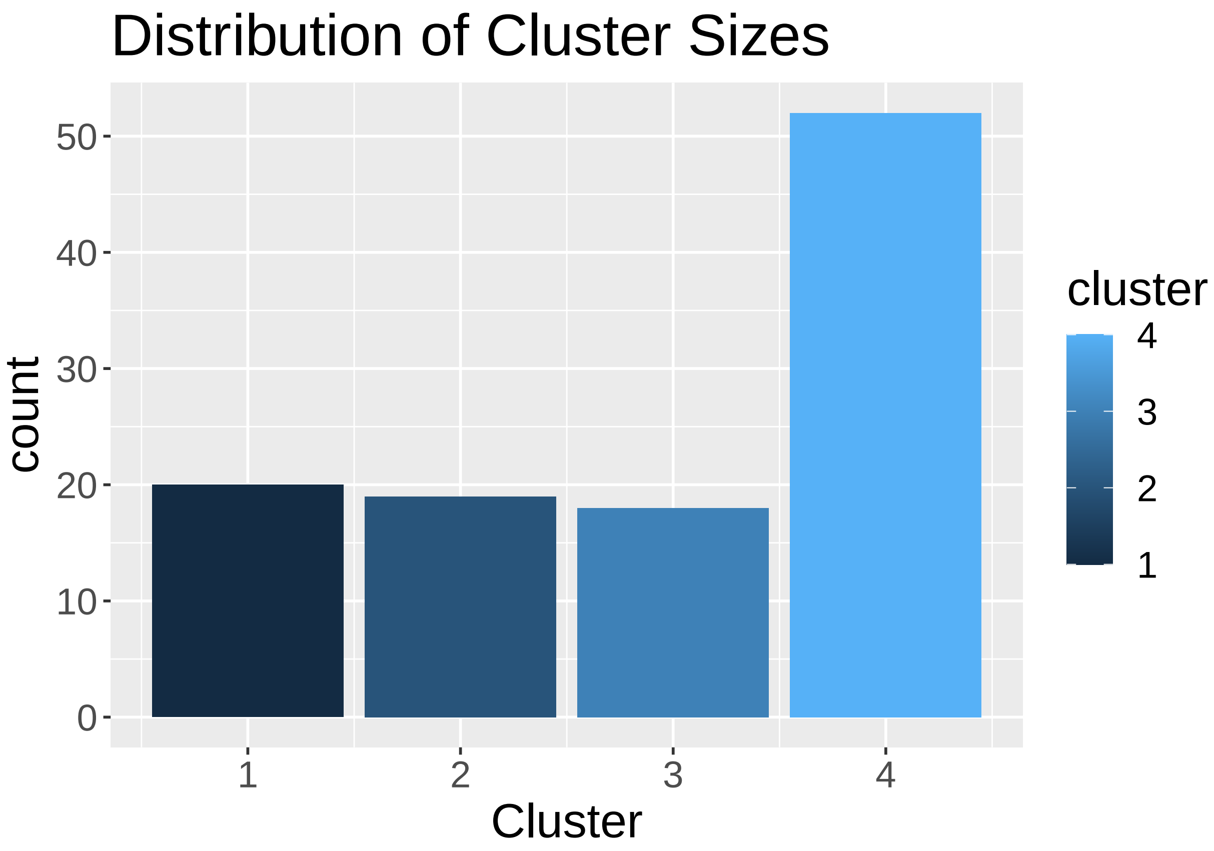


In our approach to segment this dataset into distinct groups, we're tasked with identifying the ideal number of clusters that best capture the essence of our data. This involves the utilization of the Mean Square Error (MSE) as a critical measure to determine the most representative cluster configuration, ensuring that our data segmentation is both accurate and meaningful. Based on the Elbow Method, we determined the ideal number to be 4.

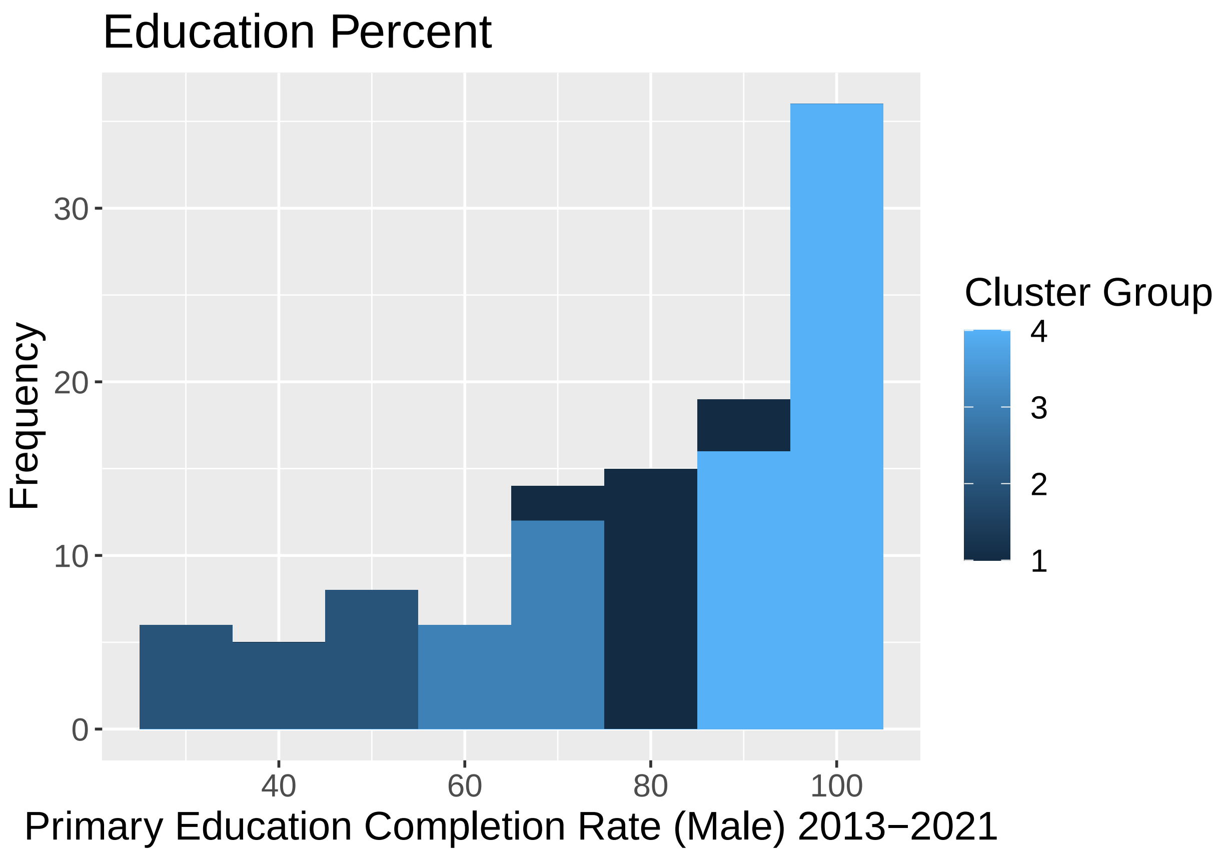
Step 4

K-Means Clustering

```r
# Step 4: K-Means Clustering
set.seed(1) # For reproducibility
k <- 4 # Optimal number of clusters found from Step 3
clustering <- kmeans(filtered_data$`sowc_education__completion_completion-rate-2013-2021-r_primary-educa

# add clustering values to our original dataset
filtered_data <-
  filtered_data %>%
  mutate(cluster = clustering$cluster)

# Number of each cluster
filtered_data %>%
  ggplot() +
  geom_bar(aes(x=cluster, group=cluster, fill=cluster)) +
  labs(x="Cluster",
      title="Distribution of Cluster Sizes") +
  theme(text=element_text(size=18))
```



```
# Plot distribution of education (color)
filtered_data %>%
  ggplot(aes(x=`sowc_education__completion_completion-rate-2013-2021-r_primary-education_male`, group=cl
  geom_histogram(binwidth=10) +
  labs(
    title="Education Percent",
    x="Primary Education Completion Rate (Male) 2013-2021",
    y="Frequency",
    fill="Cluster Group" # Label for the legend
  ) +
  theme(text=element_text(size=15))
```



Finding the optimal number of clusters to be 4, we ran the kmeans() function. We provided these arbitrary labels to the following 4 clusters. 4: High Proficiency 1: Moderate Proficiency 3: Low Proficiency 2: Extremely Low Proficiency Displayed above are distribution displaying the count of countries in each cluster. While the second distribution shows the Percentage of education completion, separated by clusters. This allows us to view which percentage is correlated to which cluster.

Step 5

Load and Filter SDG Scocres

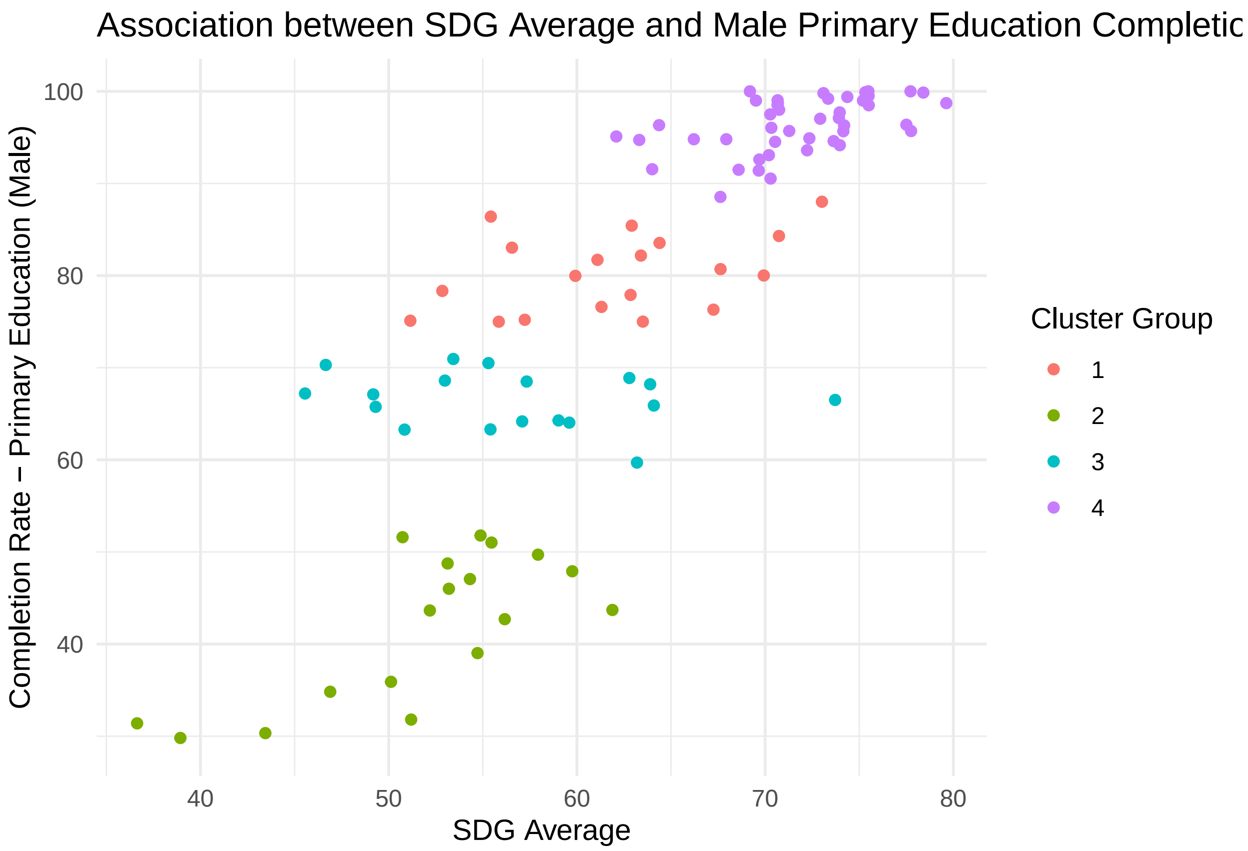
```
sdg <-
read_csv("sdr_fd5e4b5a.csv") %>%
select(-...1) # remove first column
# rename columns
names(sdg)[1:(2*17)] <-
paste(c(rep(paste("goal_", 1:17, sep=""), each=2)),
rep(c("_status", "_trend"), times=17), sep="")
names(sdg)[(2*17 + 1):(3*17)] <-
paste("goal_", 1:17, "_score", sep="")
names(sdg)[names(sdg)=="2023 SDG Index Score"] <-
"SDG_index_score_2023"
names(sdg)[names(sdg)=="2023 SDG Index Rank"] <-
"SDG_index_rank_2023"
names(sdg)[names(sdg)=="Percentage missing values"] <-
"percentage_missing_values"
```

```
names(sdg)[names(sdg)=="International Spillovers Score (0-100)"] <-
"international_spillover_score"
names(sdg)[names(sdg)=="International Spillovers Rank"] <-
"international_spillover_rank"
names(sdg)[names(sdg)=="Country Code ISO3"] <-
"country_code_iso3"

# Data joining to add goal 2 score to main data
sdg_cleaned = sdg %>% select(goal_1_score, goal_2_score, goal_3_score, goal_4_score, goal_5_score, goal_
  mutate(SDG_Average = rowMeans(select(., starts_with("goal")), na.rm = TRUE))

# join tables to add goal_2 score
data <- inner_join(x=filtered_data, y=sdg_cleaned, by="country_code_iso3")
```

Step 6

Plot the relationship between Primary Education and SDG Average

```
ggplot(data, aes(x = SDG_Average, y = data$'sowc_education__completion_completion-rate-2013-2021-r_prima
  geom_point() +  # Add points
  labs(x = "SDG Average", y = "Completion Rate - Primary Education (Male)", color = "Cluster Group") +
  theme_minimal() +  # Use a minimal theme
  ggtitle("Association between SDG Average and Male Primary Education Completion Rate")
```



Upon introducing a new column SDG_Average, we proceeded to examine the relationship between the SDG Average and the Completion Rate through plotting. It's important to note that this exploration does not constitute formal statistical evidence. We can currently see a strong correlation between the two.

Step 7

Two-Sample Hypothesis test

Null Hypothesis

$$H_0 : \mu_{SDG-I} = \mu_{SDG-J}$$

Alternative Hypothesis

$$H_1 : \mu_{SDG-I} \neq \mu_{SDG-J}$$

We choose an alpha level of 0.05.

```r
set.seed(100)
t_test_results <- list()

# Get a vector of all unique clusters
clusters <- unique(data$cluster)

# Perform t-tests between each pair of clusters and store the results
for (i in 1:(length(clusters)-1)) {
  for (j in (i+1):length(clusters)) {
    cluster_i_data <- filter(data, cluster == clusters[i])$SDG_Average
    cluster_j_data <- filter(data, cluster == clusters[j])$SDG_Average
    t_test_results[[paste("Cluster", clusters[i], "vs Cluster", clusters[j])]] <- t.test(cluster_i_data
  }
}

# Print out all the p-values
for (test_name in names(t_test_results)) {
  print(paste(test_name, "p-value:", t_test_results[[test_name]]$p.value))
}
```

```
## [1] "Cluster 3 vs Cluster 4 p-value: 3.30777405284637e-08"
## [1] "Cluster 3 vs Cluster 1 p-value: 0.0216614927527665"
## [1] "Cluster 3 vs Cluster 2 p-value: 0.0429510735883274"
## [1] "Cluster 4 vs Cluster 1 p-value: 9.74669004564009e-07"
## [1] "Cluster 4 vs Cluster 2 p-value: 4.10235905499655e-11"
## [1] "Cluster 1 vs Cluster 2 p-value: 2.73458034144235e-05"
```

By running the 2-sample hypothesis test, we can see strong evidence to reject the null hypothesis. Based on this result, we proved that the mean values of each cluster is different. And thus education is an important property that UNICEF should focus on.

Research Question 2

Is the proportion of countries that are "on track" toward achieving their SGDs by 2030 the same within South America and outside of South America?

On track: We define a country to be "on track" toward their SDGs if at least 9 out of the 17 SDGs have a goal score greater than or equal to 75% and no SDGs have a goal score below 30%.

```
# Load in the data
SDG_data <- read_csv("sdr_fd5e4b5a.csv") %>% select(-...1)  # remove first column
```

```
# Rename columns
names(SDG_data)[1:(2*17)] <-
  paste(c(rep(paste("goal_", 1:17, sep=""), each=2)),
        rep(c("_status", "_trend"), times=17), sep="")
names(SDG_data)[(2*17 + 1):(3*17)] <-
  paste("goal_", 1:17, "_score", sep="")
names(SDG_data)[names(SDG_data)=="2023 SDG Index Score"] <-
  "SDG_index_score_2023"
names(SDG_data)[names(SDG_data)=="2023 SDG Index Rank"] <-
  "SDG_index_rank_2023"
names(SDG_data)[names(SDG_data)=="Percentage missing values"] <-
  "percentage_missing_values"
names(SDG_data)[names(SDG_data)=="International Spillovers Score (0-100)"] <-
  "international_spillover_score"
names(SDG_data)[names(SDG_data)=="International Spillovers Rank"] <-
  "international_spillover_rank"
names(SDG_data)[names(SDG_data)=="Country Code ISO3"] <-
  "country_code_iso3"
```

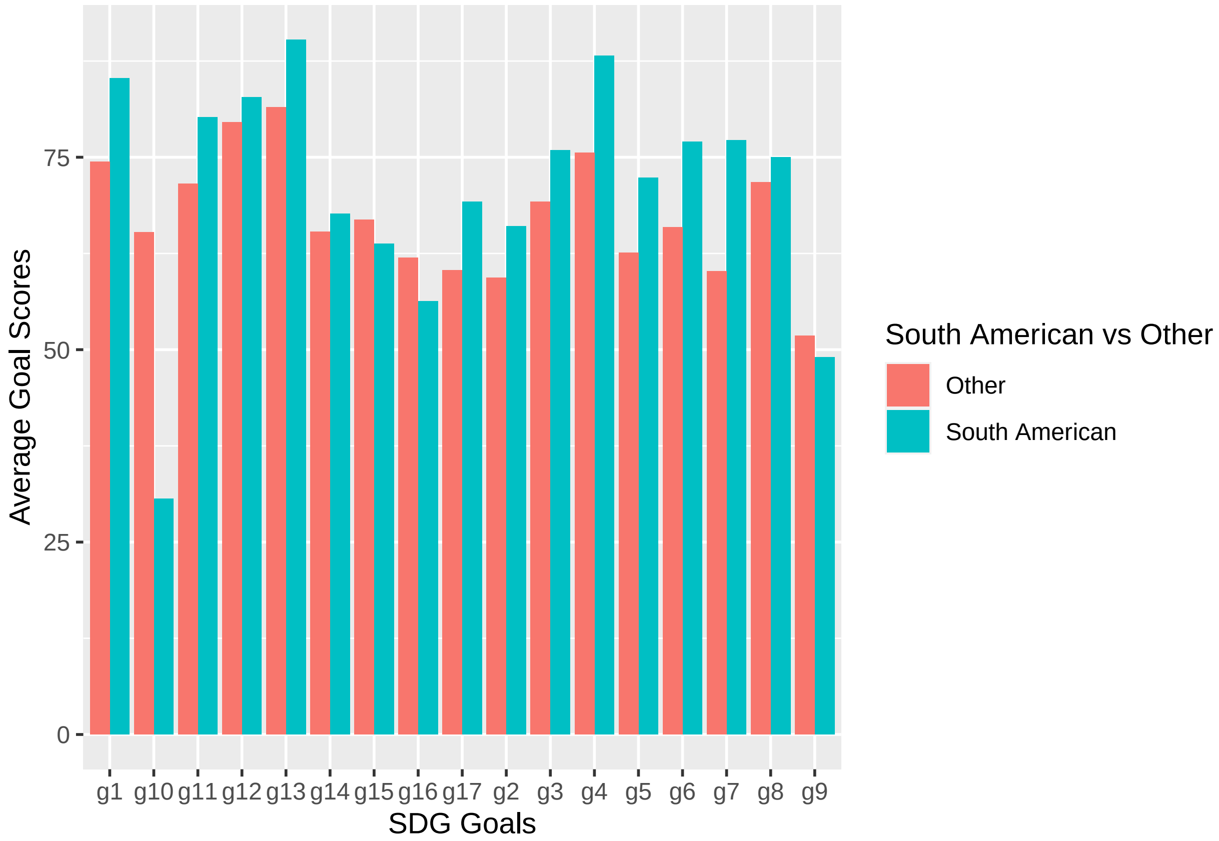
```
# Data Cleaning

# (1) Only keep the columns that are needed
SDG_data_cleaned <- SDG_data %>%
  select(goal_1_score, goal_2_score, goal_3_score, goal_4_score, goal_5_score,
         goal_6_score, goal_7_score, goal_8_score, goal_9_score, goal_10_score,
         goal_11_score, goal_12_score, goal_13_score, goal_14_score, goal_15_score,
         goal_16_score, goal_17_score, country_label)

# (2) Remove rows with NA value for country name
SDG_data_cleaned <- SDG_data_cleaned[!is.na(SDG_data_cleaned$country_label), ]


# (3) Replace missing score values with -1.
SDG_data_cleaned[is.na(SDG_data_cleaned)] <- -1


# (4) Remove the rows where there is no information for the goal scores
SDG_data_cleaned <- SDG_data_cleaned %>%
  mutate(sum = rowSums(across(where(is.numeric)))) %>%
  filter(sum != -17) %>%
  select(-sum)
```

```
# Add a new column to SDG_data_cleaned, called 'in_SA', that stores
# whether a country is in South America or not.

SA_countries = c("Brazil", "Colombia", "Argentina", "Peru", "Venezuela", "Chile",
                 "Ecuador", "Bolivia", "Paraguay", "Uruguay", "Guyana", "Suriname")

SDG_data_cleaned <- SDG_data_cleaned %>%
  mutate(in_SA = case_when(country_label %in% SA_countries ~ "Yes",
                           !(country_label %in% SA_countries) ~ "No"))
```

```
# Add a new column to SDG_data_cleaned, called 'on_track', that stores
# whether a country is on-track toward their SDG goals based on the
# following definition of on-track: A country is on-track toward their
# SDGs if at least 9 out of the 17 SDGs have a goal score greater than
# or equal to 75% and no SDGs below 30%.

SDG_data_cleaned <- SDG_data_cleaned %>% mutate(on_track = NA)
num_countries <- SDG_data_cleaned %>% summarize(n = n()) %>% as.numeric()

for (i in 1:num_countries){
  scores <- as.vector(SDG_data_cleaned[i,1:17])     # Get the ith row
  greater_75 <- length(scores[scores>75])           # Count the number of values greater than 75
  less_30 <- length(scores[scores>=0 & scores<30])  # Count the number of values less than 30

  if (greater_75 >= 9 & less_30 == 0){
    SDG_data_cleaned[i, 20] <- "Yes"
  }
  else {
    SDG_data_cleaned[i, 20] <- "No"
  }
}
```

```
# Visualization 1: Grouped Bar Chart comparing the average goal scores between
# South American and non-South American countries.

goals <- rep(c("g1", "g2", "g3", "g4", "g5", "g6",
           "g7", "g8", "g9", "g10", "g11", "g12",
           "g13", "g14", "g15", "g16", "g17"), times=2)

visual_data <- SDG_data %>% select(goal_1_score, goal_2_score, goal_3_score, goal_4_score, goal_5_score
         goal_6_score, goal_7_score, goal_8_score, goal_9_score, goal_10_score,
         goal_11_score, goal_12_score, goal_13_score, goal_14_score, goal_15_score,
         goal_16_score, goal_17_score, country_label)
visual_data <- visual_data[!is.na(visual_data$country_label), ] %>%
  mutate(in_SA = case_when(country_label %in% SA_countries ~ "Yes",
                           !(country_label %in% SA_countries) ~ "No"))

SA_countries_avg_scores <- c(visual_data %>% filter(in_SA == "Yes") %>%
                               select(-country_label, -in_SA) %>% colMeans(na.rm = TRUE))
other_countries_avg_scores <- c(visual_data %>% filter(in_SA == "No") %>%
                               select(-country_label, -in_SA) %>% colMeans(na.rm = TRUE))

SA_or_other <- rep(c("South American", "Other"), each=17)
```

```
ggplot(tibble(goals = goals, scores = c(SA_countries_avg_scores, other_countries_avg_scores), SA_or_othe
       aes(x = goals, y = scores, fill = SA_or_other)) +
  geom_bar(position = 'dodge', stat = 'identity') +
  labs(x = "SDG Goals", y = "Average Goal Scores", fill = "South American vs Other")
```



From the visualization above, we can see that South American countries have a slightly higher score for most SDG goals compared to other countries. However, South American countries appear to be struggling with goal 10 (Reduced Inequalities), since their average score for goal 10 is much lower compared to other countries. Therefore, UNICEF should support South American countries in progressing toward their SDG 10.

Two Sample Hypothesis Analysis

1. Define Sample and Compute Test Statistic

Sample: 12 South American countries vs 154 non-South American countries.

Test Statistic: Difference in proportion of countries that are on-track toward their SDGs.

```r
# Compute Test Statistic
test_stat <- SDG_data_cleaned %>% group_by(in_SA) %>%
  summarize(prop = sum(on_track == "Yes") / n(), .groups="drop") %>%
  summarize(value = abs(diff(prop))) %>% as.numeric()

test_stat
```

```
## [1] 0.1060606
```

2. Define Populations and Hypotheses

Populations: All South American countries and all non-South American countries.

Parameters: Proportion of countries that are on-track toward their SDGs for each group.

Let p_1 be the proportion of South American countries that are on-track toward their SDGs. Let p_2 be the proportion of non-South American countries that are on-track toward their SDGs.

Null Hypothesis H_0: $p_1 - p_2 = 0$

Alternative Hypothesis: H_1: $p_1 - p_2 \neq 0$

3. Simulate the Sampling Distribution

```r
set.seed(100)
num_trials <- 1000
simulated_values <- rep(NA, num_trials)
for (i in 1:num_trials){
  # Perform a random permutation
  simdata <-
    SDG_data_cleaned %>%
    mutate(in_SA = sample(in_SA, replace=FALSE))

  # Compute the simulated test statistic
  sim_test_stat <- simdata %>% group_by(in_SA) %>%
    summarize(prop = sum(on_track == "Yes") / n(), .groups="drop") %>%
    summarize(value = abs(diff(prop))) %>% as.numeric()

  simulated_values[i] <- sim_test_stat
}
```

4. Compute the p-value

We choose an alpha level of $\alpha = 0.05$. We reject the null hypothesis in favour of the alternative if the p-value is less than alpha.

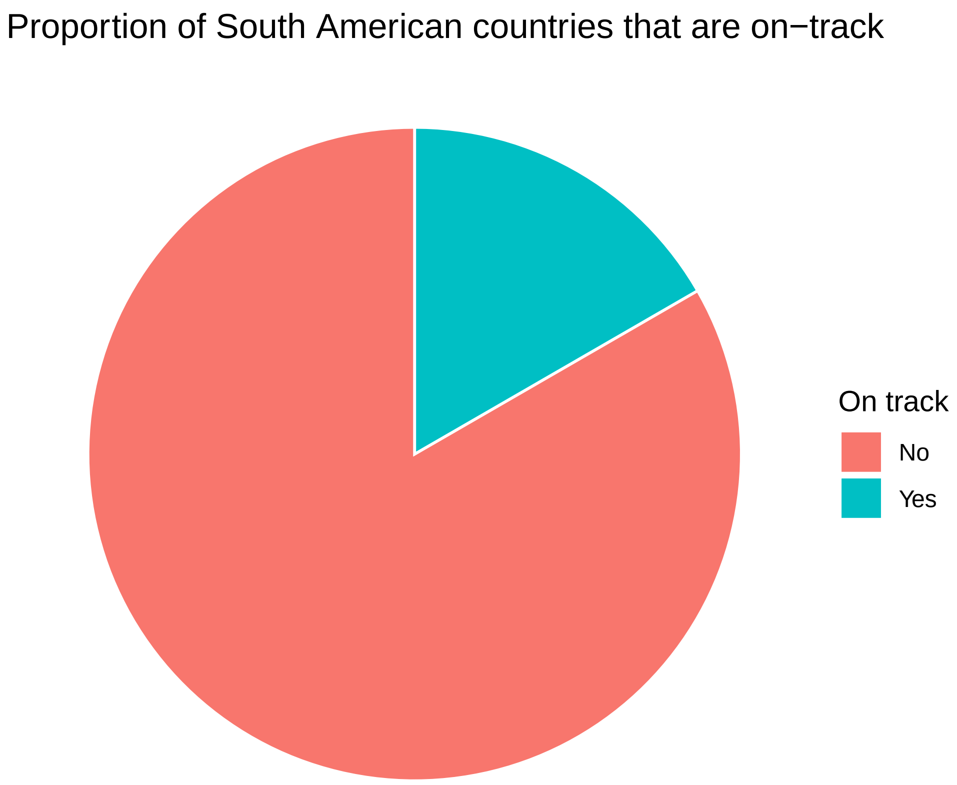
```
p_value <- sum(abs(simulated_values - 0) >= abs(test_stat - 0)) / num_trials
p_value
```

```
## [1] 0.508
```

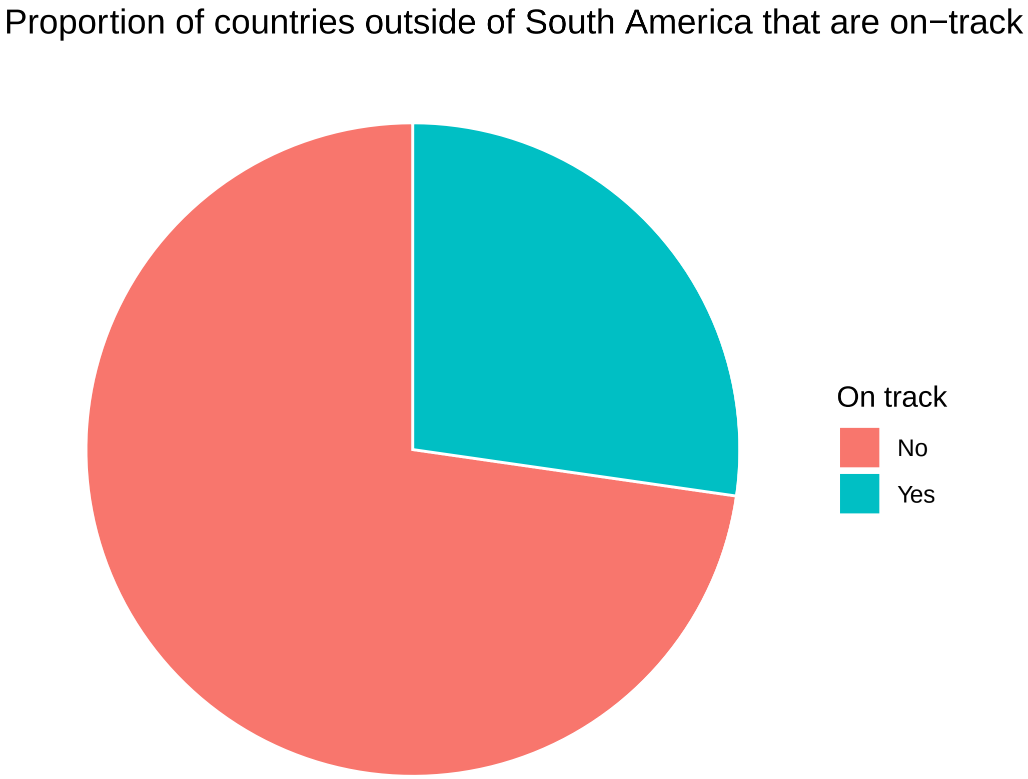
5. Make a decision and interpret

Since the p-value ($p = 0.508$) is much greater than alpha, there is no evidence against the null hypothesis. It means the proportion of countries that are on-track toward their SDGs is the same within South America compared to outside South America.

```
# Visualization 2: Pie charts showing the proportion of South American countries
# that are on-track vs the proportion of non-South American countries that are on-track.

# Pie chart 1: proportion of South American countries that are on-track
SDG_data_cleaned %>%
  filter(in_SA == "Yes") %>%
  group_by(on_track) %>%
  summarize(n=n()) %>%
  ggplot(aes(x="", y=n, fill=on_track)) +
  geom_bar(stat="identity", width=1, color="white") +
  coord_polar("y", start=0) +
  theme_void() +
  labs(fill = "On track",
       title = "Proportion of South American countries that are on-track")
```



```
# Pie chart 2: proportion of countries outside of South America that are on-track
SDG_data_cleaned %>%
  filter(in_SA == "No") %>%
  group_by(on_track) %>%
  summarize(n=n()) %>%
  ggplot(aes(x="", y=n, fill=on_track)) +
  geom_bar(stat="identity", width=1, color="white") +
  coord_polar("y", start=0) +
  theme_void() +
  labs(fill = "On track",
       title = "Proportion of countries outside of South America that are on-track")
```



From the visualization above, we can see that the proportion of countries that are on-track in South America is smaller than outside South America, which contradicts the results of our analysis. However, this difference in proportions can be explained by the large difference in group sizes. There are only 12 South American countries, whereas there are 154 non-South American countries. Therefore, each South American country has a greater effect on the proportion compared to each non-South American country. If we had one more South American country that was on-track, then the proportions would be approximately the same for both groups. Therefore, the difference is caused by only one South American country not being on-track, which is insignificant. Hence, this difference in proportions does not refute the results of our analysis.

Further Analysis

Is it possible to reject the null hypothesis with the relatively small group size of South American countries? Suppose 0 South American countries were on track. Can we reject the null hypothesis in that case?

```r
SA_countries_prop <- 0 / 12
non_SA_countries_prop <- SDG_data_cleaned %>% filter(in_SA == "No") %>%
  summarize(prop = sum(on_track == "Yes") / n(), .groups="drop") %>% as.numeric()

test_stat = abs(SA_countries_prop - non_SA_countries_prop)

p_value <- sum(abs(simulated_values - 0) >= abs(test_stat - 0)) / num_trials
p_value
```

```
## [1] 0.038
```

We can see that in the case where 0 South American countries are on track, we are able to reject the null hypothesis since $p = 0.038 < \alpha$. This shows that it is possible to reject the null hypothesis even with the relatively small group size of South American countries, however, we need a significant difference in proportions. In our analysis, we only had a difference of one South American country that was not on-track, which is not significant. Therefore, we were unable to reject the null hypothesis.

Research Question 3

What factors of South American countries are correlated with their rate of progress towards SDG 2: Zero Hunger?

```
# Load in csv files

country_indicators <-
  read_csv("country_indicators.csv") %>%
  select(-...1) %>%
  rename(country_code_iso3 = iso3)  # rename column to country_code_iso3

sdg <-
  read_csv("sdr_fd5e4b5a.csv") %>%
  select(-...1)  # remove first column

# rename columns
names(sdg)[1:(2*17)] <-
  paste(c(rep(paste("goal_", 1:17, sep=""), each=2)),
        rep(c("_status", "_trend"), times=17), sep="")
names(sdg)[(2*17 + 1):(3*17)] <-
  paste("goal_", 1:17, "_score", sep="")
names(sdg)[names(sdg)=="2023 SDG Index Score"] <-
  "SDG_index_score_2023"
names(sdg)[names(sdg)=="2023 SDG Index Rank"] <-
  "SDG_index_rank_2023"
names(sdg)[names(sdg)=="Percentage missing values"] <-
  "percentage_missing_values"
names(sdg)[names(sdg)=="International Spillovers Score (0-100)"] <-
  "international_spillover_score"
names(sdg)[names(sdg)=="International Spillovers Rank"] <-
  "international_spillover_rank"
names(sdg)[names(sdg)=="Country Code ISO3"] <-
  "country_code_iso3"
```

```
# Data joining to add goal 2 score to main data

sdg_cleaned = sdg %>% select(goal_2_score, country_code_iso3)

# join tables to add goal_2 score
data <- inner_join(x=country_indicators, y=sdg_cleaned, by="country_code_iso3")

# Remove rows with NA value for sdg goal 2
data <- data[!is.na(data$goal_2_score), ]
```

```
# Add a new column to data called 'in_SA' that stores
# whether a country is in South America or not.
SA_countries = c("BRA", "COL", "ARG", "PER", "VEN", "CHL", "ECU", "BOL", "PRY", "URY", "GUY", "SUR")

data <- data %>%
mutate(in_SA = case_when(country_code_iso3 %in% SA_countries ~ "Yes",
!(country_code_iso3 %in% SA_countries) ~ "No"))
```

```
# reorder the data to have country codes and goal_2_scores first
sa_data <- data %>% filter(in_SA == "Yes")
sa_data <- sa_data %>% select(goal_2_score, everything())
sa_data <- sa_data %>% select(country_code_iso3, everything())
```

```
predictors <- c("sowc_nutrition-newborns-preschool-school-age-children-women-and-households__weight-at-
                "sowc_nutrition-newborns-preschool-school-age-children-women-and-households__malnutriti
                "sowc_nutrition-newborns-preschool-school-age-children-women-and-households__malnutriti
                "sowc_nutrition-newborns-preschool-school-age-children-women-and-households__malnutriti
                "sowc_nutrition-newborns-preschool-school-age-children-women-and-households__malnutriti
                "sowc_nutrition-newborns-preschool-school-age-children-women-and-households__vitamin-a-
                "sowc_nutrition-newborns-preschool-school-age-children-women-and-households__malnutriti
                "sowc_nutrition-newborns-preschool-school-age-children-women-and-households__malnutriti
                "sowc_nutrition-newborns-preschool-school-age-children-women-and-households__malnutriti
                "sowc_nutrition-newborns-preschool-school-age-children-women-and-households__malnutriti
                "sowc_nutrition-newborns-preschool-school-age-children-women-and-households__percentage
                "sowc_table-9-nutrition-breastfeeding-and-diets__infant-and-young-child-feeding-0-23-mo
                "sowc_table-9-nutrition-breastfeeding-and-diets__infant-and-young-child-feeding-0-23-mo
                "sowc_table-9-nutrition-breastfeeding-and-diets__infant-and-young-child-feeding-0-23-mo
                "sowc_table-9-nutrition-breastfeeding-and-diets__infant-and-young-child-feeding-0-23-mo
                "sowc_table-9-nutrition-breastfeeding-and-diets__infant-and-young-child-feeding-0-23-mo
                "sowc_table-9-nutrition-breastfeeding-and-diets__infant-and-young-child-feeding-0-23-mo
                "sowc_table-9-nutrition-breastfeeding-and-diets__infant-and-young-child-feeding-0-23-mo
                "sowc_table-9-nutrition-breastfeeding-and-diets__infant-and-young-child-feeding-0-23-mo
                "sowc_table-9-nutrition-breastfeeding-and-diets__infant-and-young-child-feeding-0-23-mo
                "sowc_table-9-nutrition-breastfeeding-and-diets__infant-and-young-child-feeding-0-23-mo
                "sowc_table-9-nutrition-breastfeeding-and-diets__infant-and-young-child-feeding-0-23-mo
                "sowc_table-9-nutrition-breastfeeding-and-diets__infant-and-young-child-feeding-0-23-mo
```

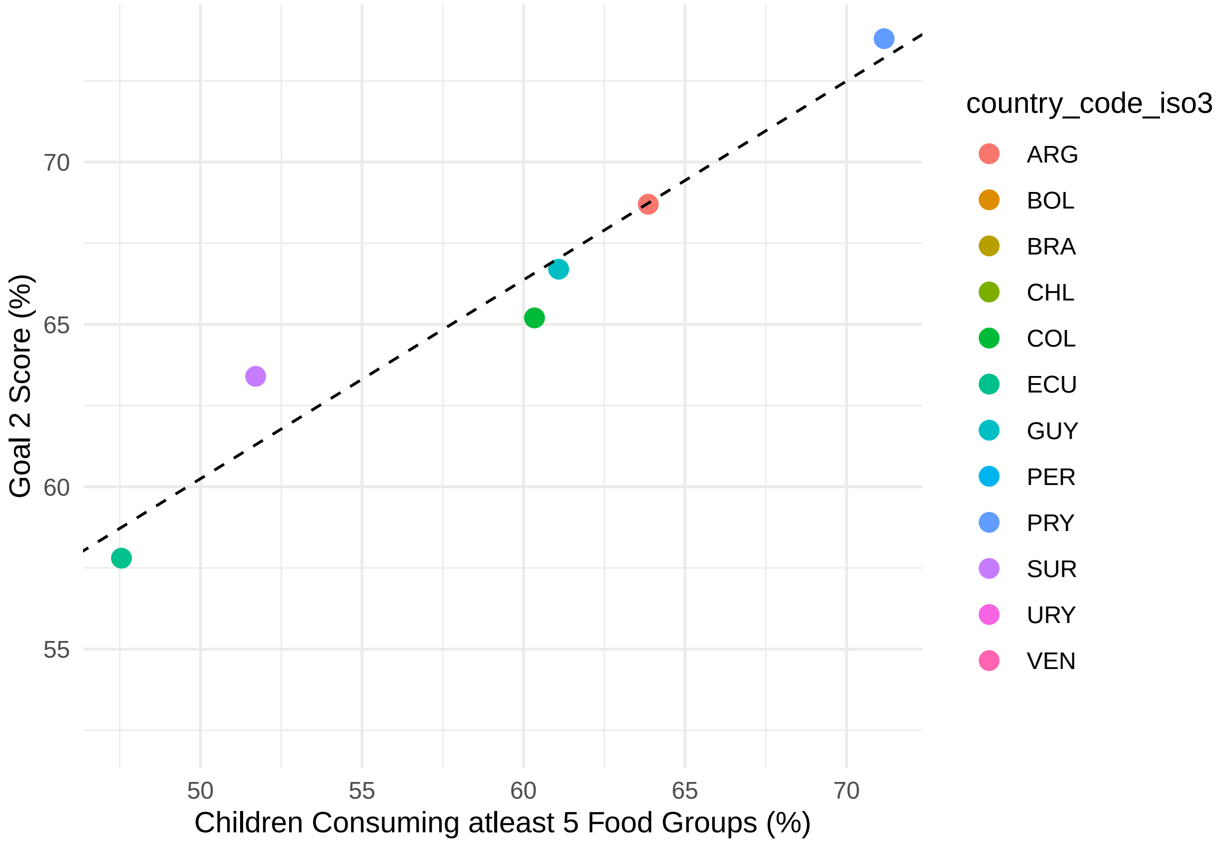
```
set.seed(100)
# compute the regression models
regression_models <- list()

# Fit simple linear regression for each predictor
for (predictor in predictors) {
  # Create formula for regression
  formula <- as.formula(paste("goal_2_score ~ `", predictor, "`", sep = ""))
  predictor_coloumn <- as.name(predictor)

  # filter the data to remove NA values
  filtered_data <- sa_data %>% filter(!is.na(!!sym(predictor)))

  # Fit the linear regression model
  # Check if there is data to fit on
  if (nrow(filtered_data) > 0) {
    # Fit the linear regression model
    model <- lm(formula, data = filtered_data)

    # Store the model in the list
    regression_models[[predictor]] <- model
  }
}
```

```
# Get the indices of the models sorted by R-squared values
# used ChatGPT to order the models by R-squared values
sorted_indices <- order(sapply(regression_models, function(model) summary(model)$r.squared), decreasing

# Reorder the list of models based on the sorted indices
sorted_models <- regression_models[sorted_indices]

# print out the coefficients
for (model in sorted_models[1:3]) {
  print("model details:")
  # this will print the intercept and slope
  coefficients <- coef(model)
  print("intercept:")
  print(as.numeric(coefficients[1]))
  print("slope:")
  print(as.numeric(coefficients[2]))
  print("variable name:")
  print(names(coefficients)[2])
}
```

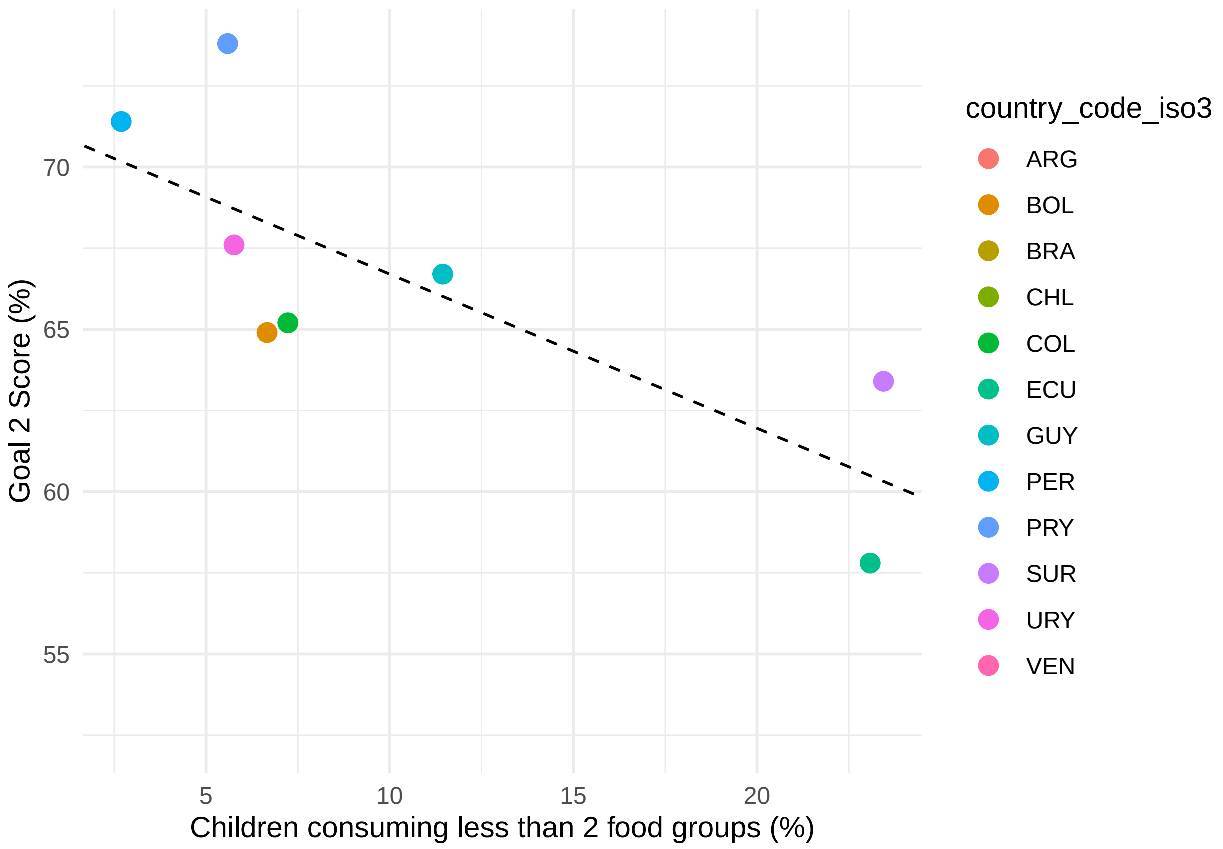
```
## [1] "model details:"
## [1] "intercept:"
## [1] 29.62928
## [1] "slope:"
## [1] 0.6123611
## [1] "variable name:"
## [1] "‘sowc_table-9-nutrition-breastfeeding-and-diets__infant-and-young-child-feeding-0-23-months-201
## [1] "model details:"
## [1] "intercept:"
## [1] 71.44735
## [1] "slope:"
## [1] -0.4747488
## [1] "variable name:"
## [1] "‘sowc_table-9-nutrition-breastfeeding-and-diets__infant-and-young-child-feeding-0-23-months-201
## [1] "model details:"
## [1] "intercept:"
## [1] 35.69062
## [1] "slope:"
## [1] 0.3647213
## [1] "variable name:"
## [1] "‘sowc_table-9-nutrition-breastfeeding-and-diets__infant-and-young-child-feeding-0-23-months-201
```

```
# Graph with lowest R-Squared
plot1 <- ggplot(sa_data, aes(x = `sowc_table-9-nutrition-breastfeeding-and-diets__infant-and-young-chil
                             y = goal_2_score,
                             colour = country_code_iso3)) +
  geom_point(size = 3) +
  geom_abline(intercept = 29.6292753, slope = 0.6123611 , linetype = "dashed") +
  labs(x = "Children Consuming atleast 5 Food Groups (%)",
       y = "Goal 2 Score (%)") +
  theme_minimal()
```

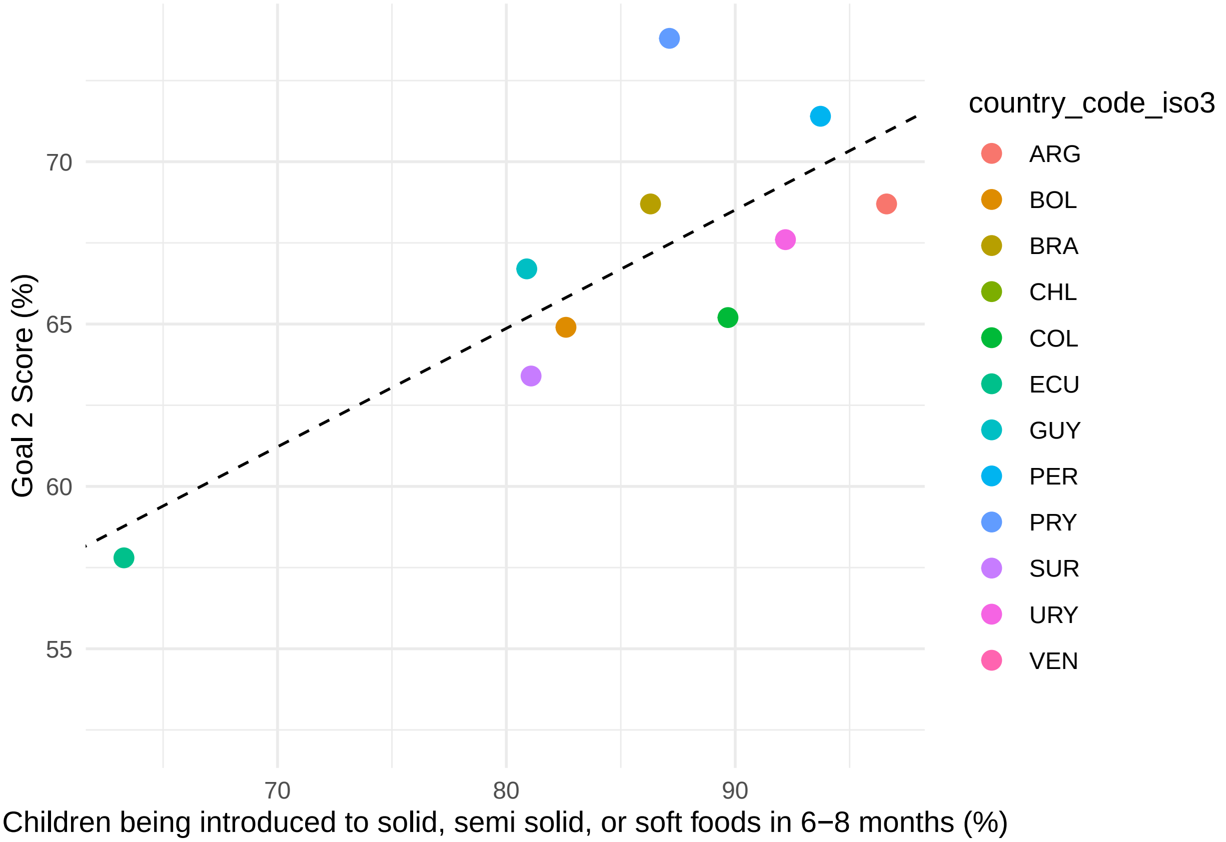
```
# Graph with second lowest R-Squared
plot2 <- ggplot(sa_data, aes(x = `sowc_table-9-nutrition-breastfeeding-and-diets__infant-and-young-chil
                    colour = country_code_iso3)) +
  geom_point(size = 3) +
  geom_abline(intercept = 71.4473, slope = -0.4747 , linetype = "dashed") +
  labs(x = "Children consuming less than 2 food groups (%)",
       y = "Goal 2 Score (%)") +
  theme_minimal()


# Graph with third lowest R-Squared
plot3 <- ggplot(sa_data, aes(x = `sowc_table-9-nutrition-breastfeeding-and-diets__infant-and-young-chil
                             y = goal_2_score,
                             colour = country_code_iso3)) +
  geom_point(size = 3) +
  geom_abline(intercept = 35.6906, slope = 0.3647  , linetype = "dashed") +
  labs(x = "Children being introduced to solid, semi solid, or soft foods in 6-8 months (%)",
       y = "Goal 2 Score (%)") +
  theme_minimal()

plot1
```



```
plot2
```


country_code_iso3
ARG
BOL
BRA
CHL
COL
ECU
GUY
PER
PRY
SUR
URY
VEN
Goal 2 Score (%)
70
65
60
55
5
10
15
20
Children consuming less than 2 food groups (%)

```
plot3
```



Visual Analysis

As we see from the above models, there is a strong connection between the percent of children consuming at least 5 food groups and the corresponding South American country's SDG 2 score. This implies a positive correlation that if more countries have greater populations with access to 5 healthy food groups, then it improves the SDG 2 score for the country since less children will be hungry and live healthier over all.

Plot 2 shows a negative correlation between the percentage of children consuming less than 2 food groups and the country's SDG score. This indicates that those countries that struggle to feed different healthy foods to children struggle in meeting SDG score two, which implies more emphasis required on providing struggling countries with a greater variety of food groups.

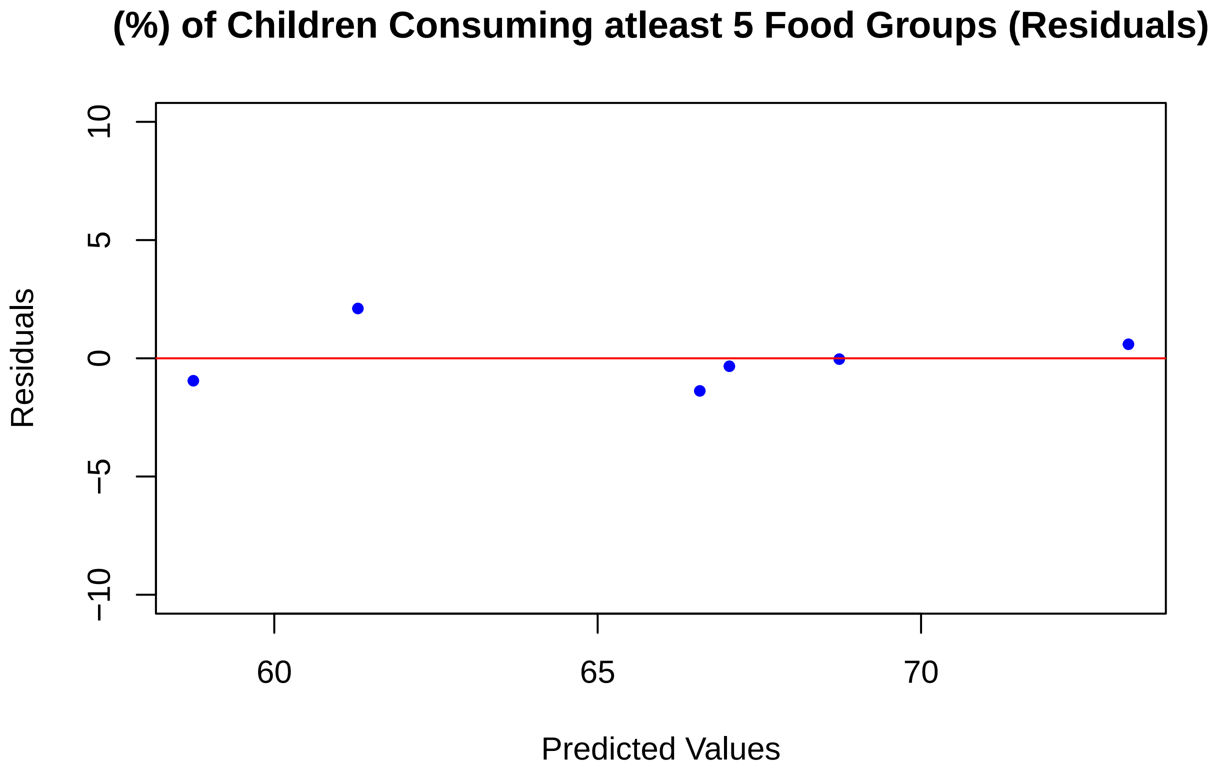
From plot 3, we can see a positive correlation between percent of children aged 6-8 months being introduced to solid or soft foods and a country's SDG score. This can lead us to infer that children that have higher nutritional intake in forms of solid foods and not just liquids implies that a country is more well suited to over come hunger.

P Value Tests on Two Models

Here we choose an alpha level of $\alpha = 0.01$. Here we define our null hypothesis $H_0 : \beta_1 = 0$ and our alternative as $H_1 : \beta_1 \neq 0$, where β_1 is our coefficient in our regression model. We reject H_0 if our computed P-value is less than our chosen α. We will apply this test to two of our chosen models.

Model 1 Residuals

```r
set.seed(100)
# Create model 1 to show the residual plot and P-Value
model1 <- lm(goal_2_score ~ `sowc_table-9-nutrition-breastfeeding-and-diets__infant-and-young-child-fee

model1_summary <- summary(model1)

# the p value for the predictor term
p_values <- model1_summary$coefficients[, "Pr(>|t|)"]

# Extract only the numeric values (p-values)
numeric_p_values <- as.numeric(p_values)

predicted_values <- predict(model1)

# Get residuals from the model
residuals <- residuals(model1)

plot(predicted_values, residuals,
     xlab = "Predicted Values",
     ylab = "Residuals",
     main = "(%) of Children Consuming atleast 5 Food Groups (Residuals)",
     pch = 20,
     col = "blue",
     ylim = range(c(-10, 10)))

abline(h = 0, col = "red")
```



```
# print the variable p-value
print(numeric_p_values[2])
```

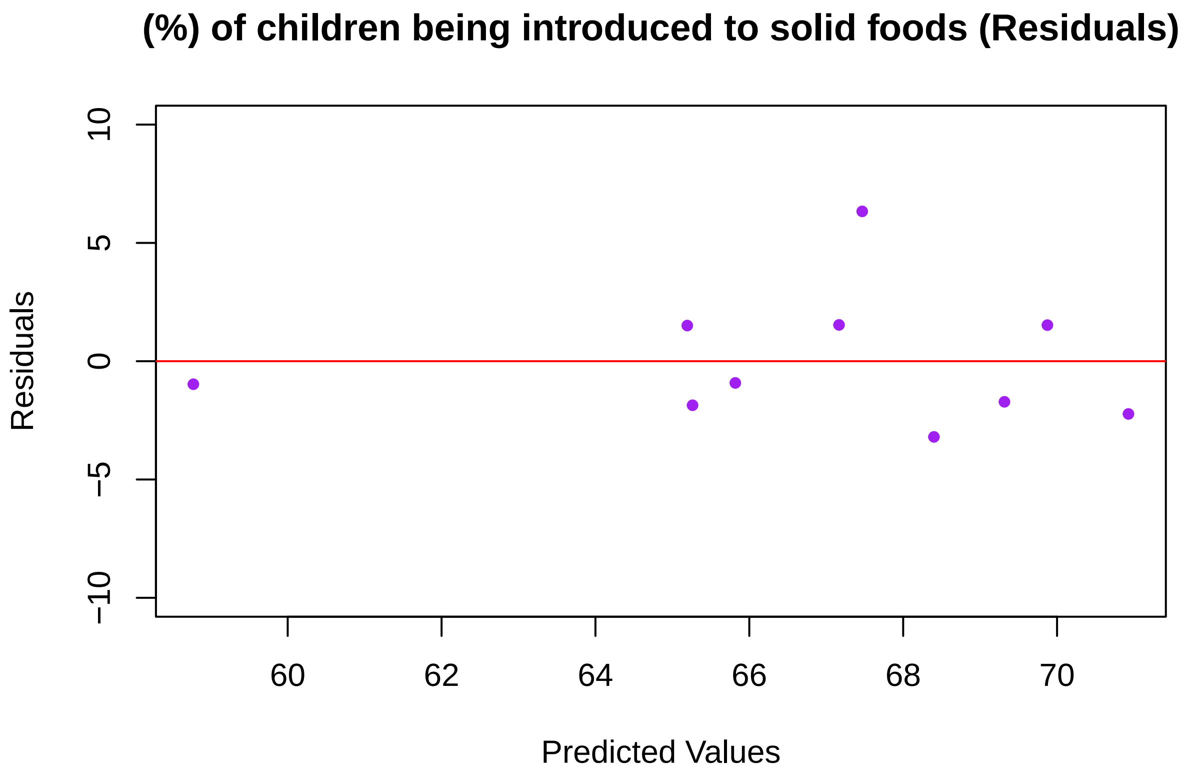
```
## [1] 0.001106231
```

Here we see that our computed p-value for β_1 is evaluated at 0.001106231 which is less than our chosen α value of 0.05, hence we can reject H_0 and conclude that the percent of children consuming at least 5 food groups has correlation with a country's SDG score. We can also visually see in the residual plot that the model is generally accurate within 2.5% of the over SDG 2 score.

Model 2 Residuals

```
set.seed(100)
# Create model 2 to show the residual plot and P-Value
model2 <- lm(goal_2_score ~ `sowc_table-9-nutrition-breastfeeding-and-diets__infant-and-young-child-fee

model2_summary <- summary(model2)
# the p value for the predictor term
p_values2 <- model2_summary$coefficients[, "Pr(>|t|)"]

# Extract only the numeric values (p-values)
numeric_p_values2 <- as.numeric(p_values2)
```

```
predicted_values2 <- predict(model2)
residuals2 <- residuals(model2)

plot(predicted_values2, residuals2,
     xlab = "Predicted Values",
     ylab = "Residuals",
     main = "(%) of children being introduced to solid foods (Residuals)",
     pch = 20,
     col = "purple",
     ylim = range(c(-10, 10)))

abline(h = 0, col = "red")
```



```
print(numeric_p_values2[2])
```

```
## [1] 0.008392281
```

For our second plot, we can see that our computed p-value for β_1 is 0.008392. This satisfies $0.008392 < \alpha$ meaning we can reject H_0 for this model as well. By assuming H_1, we can conclude that there is a correlation between the percent of children aged from 6-8 months being introduced to solid foods and a country's SDG score. This implies that more emphasis should be placed on making baby foods more accessible. Fairly accurate predictions are depicted by this model, visually represented on the residual plot showing that most of the predictions are within 5% of a country's SDG 2 score.

ChatGPT Citations:

ChatGPT Prompt:
“given a list of linear regression models, how to sort them based off of their r sqaured values in R”
ChatGPT Response:
‘# Assuming ’linear_models’ is your list of linear regression models’

‘# Extract the R-squared values from each model’ r_squared_values <- sapply(linear_models, function(model) summary(model)$r.squared)

‘# Get the indices of the models sorted by R-squared values’ sorted_indices <- order(r_squared_values, decreasing = TRUE)

‘# Reorder the list of models based on the sorted indices’ sorted_linear_models <- linear_models[sorted_indices]

In this code:

sapply() is used to apply a function (in this case, extracting the R-squared value) to each model in the list. order() sorts the indices of the models based on their R-squared values, with decreasing = TRUE sorting in descending order. Finally, the list of linear regression models is reordered based on the sorted indices. After running this code, sorted_linear_models will contain your linear regression models sorted based on their R-squared values. Adjust the variable names according to your specific use case.