STA130 Project

2024-03-27

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Background

A lot of Southern Asian countries are facing serious public health issues. By comparing the progress of South Asian countries towards SDG Goal 3 with the rest of the world, we hope to draw conclusions about how UNICEF could accelerate their progress. We will be comparing 3 aspects: the under-five child mortality rate, the association between available family planning resources and maternal mortality, & the clusters of countries with similar progress.

Research Question 1:

Is the mean of the annual rate of reduction in under-five child mortality rate similar between South Asian countries and the rest of the world?

We will first import the data

```
unicef_data <- read_csv("country_indicators.csv")
```

We filter out the Columns that we need

```
#We select the two columns we need.
child_mortality_red_rates <- unicef_data %>% select( ...1, u5mr)
```

We then need to clean the data that we have selected

```
#We remove all the rows that don't have a value.
child_mortality_red_rates2 <- child_mortality_red_rates %>%
  filter(rowSums(is.na(.)) == 0)
child_mortality_red_rates2
```

```
## # A tibble: 198 x 2
##     ...1  u5mr
##    <dbl> <dbl>
##  1     0   4
##  2     1   5
##  3     2   3
##  4     3   4.8
##  5     4   5.2
##  6     5   4.1
##  7     6   4.4
##  8     7   5
##  9     8   5
```

```
## 10     9   2.4
## # i 188 more rows
```

We will make a vector of the country codes of all South Asian countries and then filter out those values. We will then take their average, and take the average of the child mortality rate reduction for the rest of the countries and subtract them. This data will be our test statistic

```
# This vector contains the country codes of all of the South Asian countries
a <- c(179,180,181,182,183,184,185,186,187)
#We filter out South Asian Countries
south_asia <- child_mortality_red_rates2 %>% filter( ...1 %in% a)
#We filter out rest of the countries
rest_countries <- child_mortality_red_rates2 %>% filter(!(...1 %in% a))

n_south <- 3
n_world <- 185

# randomizing the sample
south_asia_average <- south_asia %>% mutate(...1 = sample(...1))

#Taking the mean of 5 values
south_asia_average1 <- slice(south_asia_average,
                             (n_south + 1):nrow(south_asia))%>%summarise ( Average =
               mean(u5mr)) %>% as.numeric()


rest_countries_average <- rest_countries %>% mutate(...1 = sample(...1)) #randomizing the sample

#Taking the mean of 5 values
rest_countries_average1 <- slice(rest_countries_average,
                                 (n_world + 1):nrow(rest_countries)) %>%
  summarise ( Average = mean(u5mr)) %>% as.numeric()

#Difference in the mean of South Asian countries from the mean of rest of the countries.
test_statistic <- rest_countries_average1 - south_asia_average1 %>% as.numeric()
```

Now we will define our null hypothesis and alternative hypothesis

1. The sample size is 5 countries from South Asia and 5 countries from the rest of the world.
2. Population: All the countries in the world
3. Parameter: Mean U5MR of each permutation.
4. Null Hypothesis (H_0): There is no significant difference between the rate of reduction of under five child mortality from 2000 - 2021 for the world and South Asian countries.
5. Alternative Hypothesis (H_1): There is a significant difference between the rate of reduction for the under five child mortality rate from 2000 - 2021 for the world and South Asian countries.
6. Alpha level(α): 0.05 to avoid Type I errors.

Since we already have the test statistic, we perform the simulation

```
set.seed(130)
# setup
num_trials <- 1000  # number of simulations/trials
mean_simulations <- rep(NA, num_trials)
for(i in 1:num_trials){
```

```
  n_south <- 3
  n_world <- 185

  south_asia_random <- south_asia %>% mutate(u5mr = sample(u5mr, replace = FALSE))
  #We randomize the sample

  south_asia_random1 <- slice(south_asia_random, (n_south + 1):nrow(south_asia)) %>%
    summarise ( Average = mean(u5mr)) %>% as.numeric()
  #Taking mean of 5 countries

  rest_countries_random <- rest_countries %>% mutate(u5mr = sample(u5mr, replace = FALSE))
  #We randomize the sample

  rest_countries_random1 <- slice(rest_countries_random,
                                  (n_world + 1):nrow(rest_countries)) %>%
    summarise ( Average = mean(u5mr)) %>% as.numeric()
  #Taking mean of 5 countries

  #Difference between the means
  difference <- rest_countries_random1 - south_asia_random1 %>% as.numeric()
  mean_simulations[i] <- difference #Storing the values

}
#making a tibble of the values
simulation_results <- tibble(mean_diff = mean_simulations)
```

Now we will find the p-value for the data and make the plots

```
delta_mean_null <- 0

pvalue_2side <-
  sum(abs(mean_simulations - delta_mean_null) >=
      abs(test_statistic - delta_mean_null)) / num_trials

print(pvalue_2side)
```

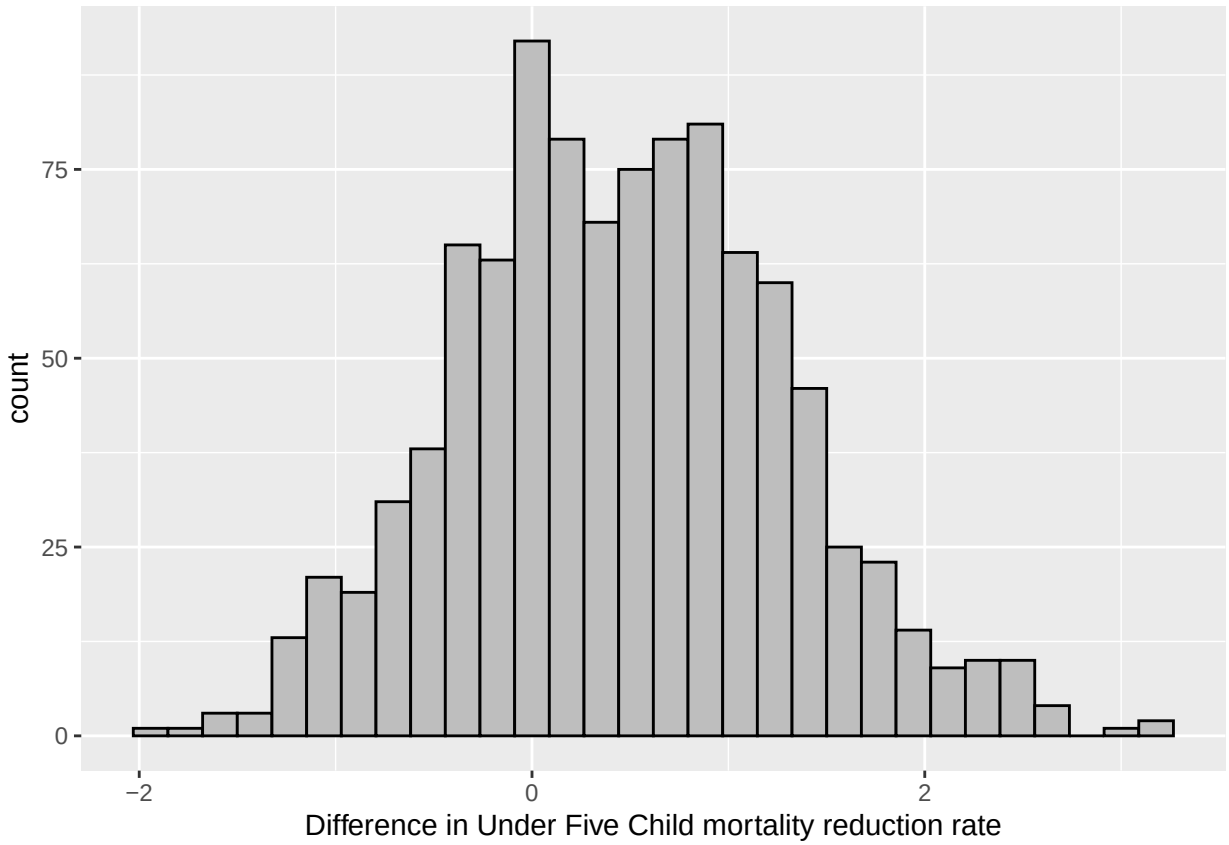
```
## [1] 0.259
```

Conclusion

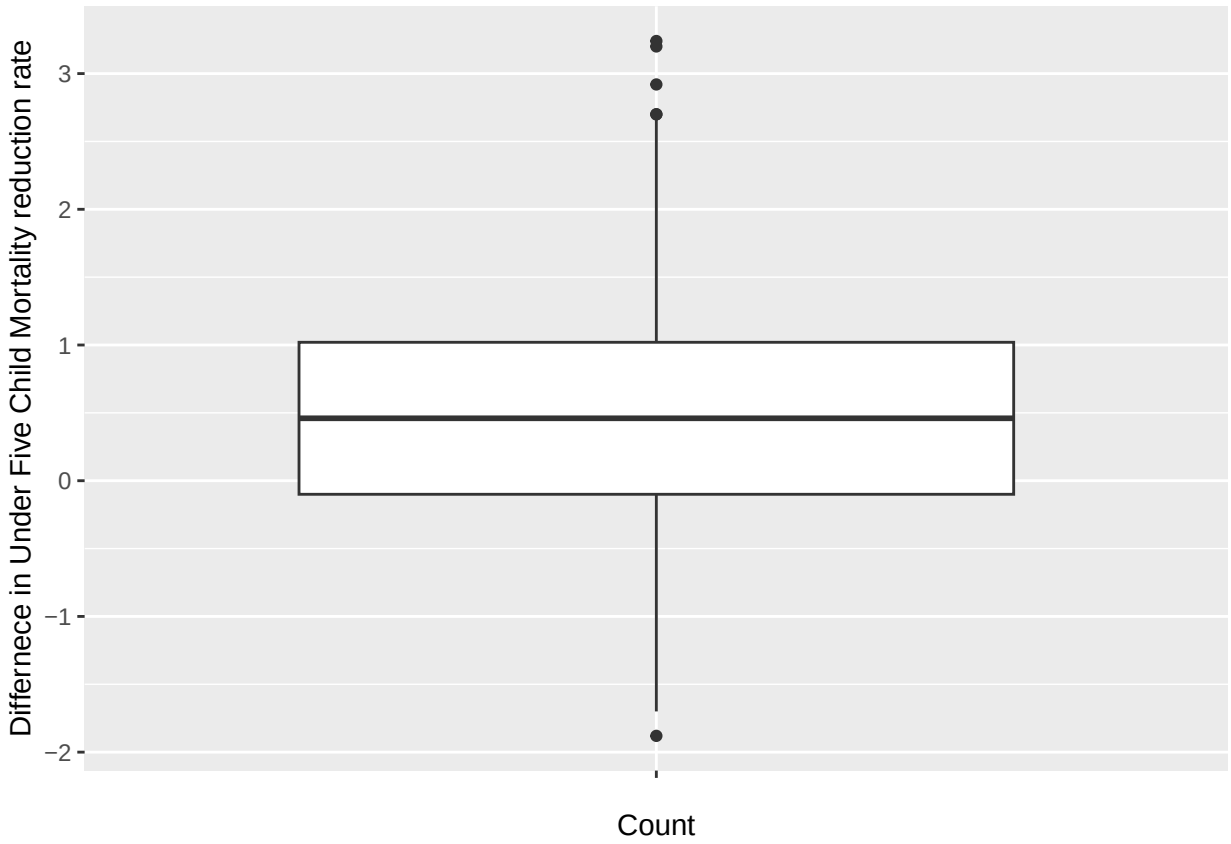
Since, the p-value is greater than 0.05, we don't have enough evidence to reject the null hypothesis. Thus, the U5MR reduction rate for South Asian countries is similar to the rest of the world.

Plots

```
  ggplot(data = simulation_results ,aes(x = mean_diff)) +
  geom_histogram(bins=30, color="black", fill="gray") +
  labs(x = "Difference in Under Five Child mortality reduction rate")
```




```
ggplot(data = simulation_results , aes(x = "",y = mean_diff)) +
geom_boxplot() +
labs(x = "Count" ,y = "Differnece in Under Five Child Mortality reduction rate")
```



Analysis

Histogram:

As we can see, the histogram is symmetrical and uni-modal having a peak at 0. This means that in most permutations South Asian countries and countries from the rest of the world have a similar U5MR reduction rate. The range of the histogram is from a little less than -2 to 3.20, with most of the observations in the middle, this also suggests that the U5MR is similar for South Asian countries and the rest of the world.

Boxplot:

From the boxplot we can see that the IQR lies between 0 and 1, and the median lies around 0.5. This means that most of the values don't have a significant difference in the U5MR reduction rate. We only have a couple of outliers in the boxplot, this also suggests that most values lie in the middle where the difference is small, which means that in this simulation most of the times there isn't a significant difference in the U5MR reduction rate of South Asia and the rest of the world.

Potential Biases

The data for some countries across the world is missing which might cause the mean of U5MR to be different from the actual value. Since this is the case for very few countries the mean we obtain will still be accurate to a high degree.

We believe that there may be potential confounding variables that can affect the observed relationship between the two variables, such as education level, sanitary practices and access to resources. A country that has better education will also have citizens who are more aware of sanitary practices and proper procedures which lowers the U5MR.

Reasearch Question 2

Is there an association between available family planning resources and maternal mortality? What does this imply for South Asian countries?

```
# Load in the country_codes CSV file.
country_codes <- read_csv("./country_codes.csv")
```

```
# Load in the Table-3-Maternal-and-newborn-health-SOWC2023 Excel Sheet.
data <- read_excel("./Table-3-Maternal-and-newborn-health-SOWC2023 (6).xlsx")
```

```
# Select the two columns Country or Area_en (M49) and Sub-region Name_en (M49)
# from the country_codes dataset, rename them, remove rows that contain missing values (NA),
# and store them in a new dataset country_indicators.
country_indicators <-
  country_codes %>%
  select(`Country or Area_en (M49)`, `Sub-region Name_en (M49)`) %>%
  rename(c("country_names" = "Country or Area_en (M49)",
           "sub_regions" = "Sub-region Name_en (M49)")) %>%
  mutate(sub_regions=if_else(sub_regions == "Southern Asia", "Southern Asia",
                             "Rest of the World")) %>%
  na.omit()

country_indicators
```

```
## # A tibble: 254 x 2
##    country_names                  sub_regions
##    <chr>                          <chr>
##  1 Algeria                        Rest of the World
##  2 Egypt                          Rest of the World
##  3 Libya                          Rest of the World
##  4 Morocco                        Rest of the World
##  5 Sudan                          Rest of the World
##  6 Tunisia                        Rest of the World
##  7 Western Sahara                 Rest of the World
##  8 British Indian Ocean Territory Rest of the World
##  9 Burundi                        Rest of the World
## 10 Comoros                        Rest of the World
## # i 244 more rows
```

```
# Select the three columns TABLE 3. MATERNAL AND NEWBORN HEALTH, ...6, and ...24 from the
# data dataset, rename them, and store them in a new dataset maternal_and_newborn_health.
# Then remove rows from the maternal_and_newborn_health dataset that contain missing values
# (NA). Using the filter() function, return a subset of rows that don't contain the dash '-';
# this effectively removes missing values ('-') as well. head(129) function returns the 1st 129
# rows of the maternal_and_newborn_health dataset.
maternal_and_newborn_health <-
  data %>%
  select(`TABLE 3. MATERNAL AND NEWBORN HEALTH`, `...6`, `...24`) %>%
  rename(c("country_names" = "TABLE 3. MATERNAL AND NEWBORN HEALTH",
      "demand_family_planning" = "...6",
      "maternal_mortality_ratio" = "...24")) %>%
  na.omit() %>%
  filter(demand_family_planning != '-' & maternal_mortality_ratio != '-') %>%
  mutate(demand_family_planning=round(as.numeric(demand_family_planning),0),
         maternal_mortality_ratio=round(as.numeric(maternal_mortality_ratio)), 0) %>%
```

```
  head(129)

maternal_and_newborn_health
```

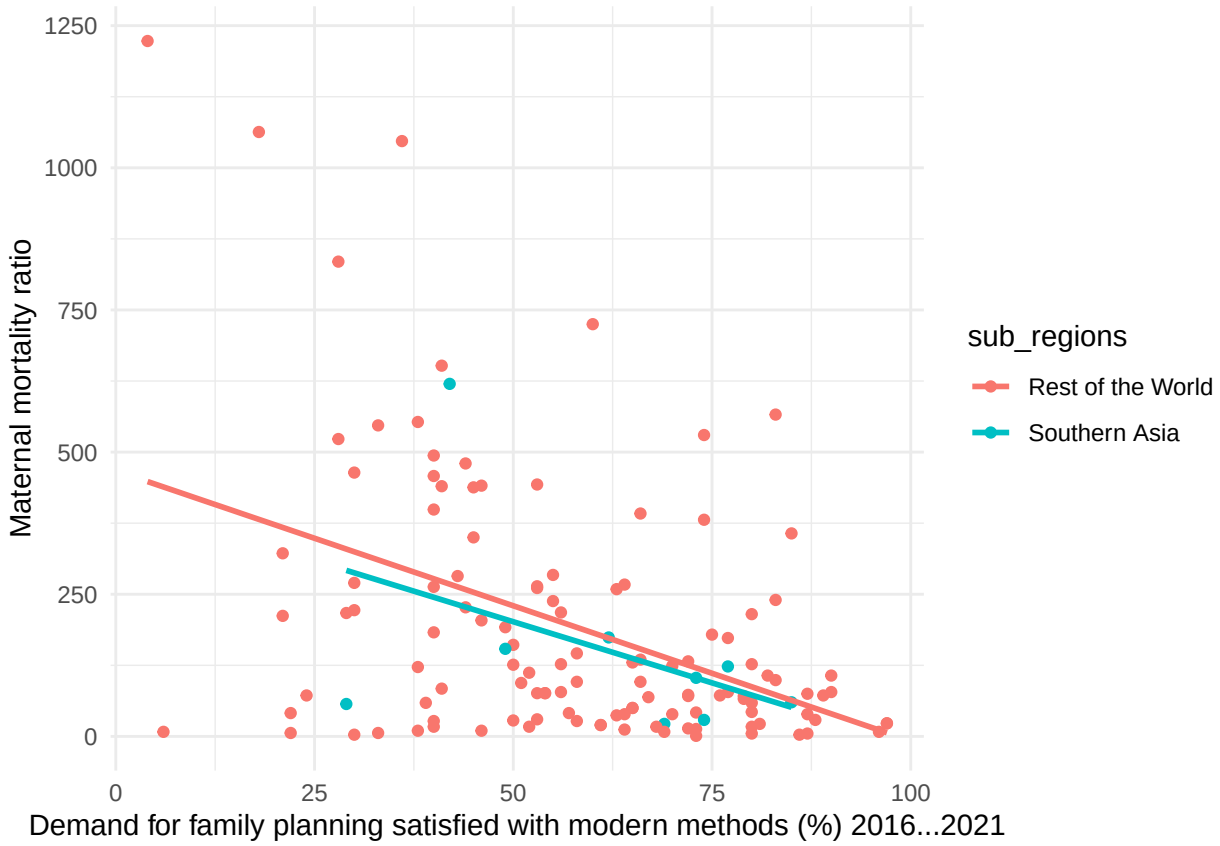
```
## # A tibble: 129 x 4
##    country_names demand_family_planning maternal_mortality_ratio   `0`
##    <chr>                          <dbl>                    <dbl> <dbl>
##  1 Afghanistan                       42                      620     0
##  2 Albania                            6                        8     0
##  3 Algeria                           77                       78     0
##  4 Angola                            30                      222     0
##  5 Armenia                           40                       27     0
##  6 Azerbaijan                        22                       41     0
##  7 Bangladesh                        77                      123     0
##  8 Barbados                          70                       39     0
##  9 Belarus                           73                        1     0
## 10 Belize                            65                      130     0
## # i 119 more rows
```

```
# Merge the country_codes CSV file with Table-3-Maternal-and-newborn-health-SOWC2023 Excel
# Sheet based on matching rows of their country_names into a single data frame.

merged_data <- inner_join(maternal_and_newborn_health, country_indicators, by="country_names")
merged_data
```

```
## # A tibble: 129 x 5
##    country_names demand_family_planning maternal_mortality_r~1   `0` sub_regions
##    <chr>                          <dbl>                  <dbl> <dbl> <chr>
##  1 Afghanistan                       42                    620     0 Southern A~
##  2 Albania                            6                      8     0 Rest of th~
##  3 Algeria                           77                     78     0 Rest of th~
##  4 Angola                            30                    222     0 Rest of th~
##  5 Armenia                           40                     27     0 Rest of th~
##  6 Azerbaijan                        22                     41     0 Rest of th~
##  7 Bangladesh                        77                    123     0 Southern A~
##  8 Barbados                          70                     39     0 Rest of th~
##  9 Belarus                           73                      1     0 Rest of th~
## 10 Belize                            65                    130     0 Rest of th~
## # i 119 more rows
## # i abbreviated name: 1: maternal_mortality_ratio
```

Including Plots

```
# Univariate Linear Regression Model -----------------

# Plot:
merged_data %>% ggplot(aes(x=demand_family_planning, y=maternal_mortality_ratio,
                           color=sub_regions)) +
  geom_point() +
  labs(x="Demand for family planning satisfied with modern methods (%) 2016-2021",
       y="Maternal mortality ratio") +
  geom_smooth(method="lm", se=FALSE) +
  theme_minimal()
```




```
# Split data into southern asia and rest of the world data sets to remove outliers from
# each set
split_data <- split(merged_data, merged_data$sub_regions)
southern_asia = split_data[["Southern Asia"]]
rest_world = split_data[["Rest of the World"]]
```

```
# SOUTHERN ASIA linear regression model + correlation value
model_southern_asia <- lm(maternal_mortality_ratio ~ demand_family_planning, data=southern_asia)

summary(model_southern_asia)
```

```
##
## Call:
## lm(formula = maternal_mortality_ratio ~ demand_family_planning,
##     data = southern_asia)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -234.98  -69.46    0.24   23.93  383.92
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             416.70     220.29   1.892    0.100
## demand_family_planning   -4.30       3.41  -1.261    0.248
##
## Residual standard error: 178 on 7 degrees of freedom
## Multiple R-squared:  0.1852, Adjusted R-squared:  0.06878
## F-statistic: 1.591 on 1 and 7 DF,  p-value: 0.2476
```

```r
cor(x=southern_asia$demand_family_planning, y=southern_asia$maternal_mortality_ratio)
```

```
## [1] -0.4303322
```

```r
# REST OF WORLD linear regression model + correlation value
model_rest_world <- lm(maternal_mortality_ratio ~ demand_family_planning, data=rest_world)

summary(model_rest_world)
```

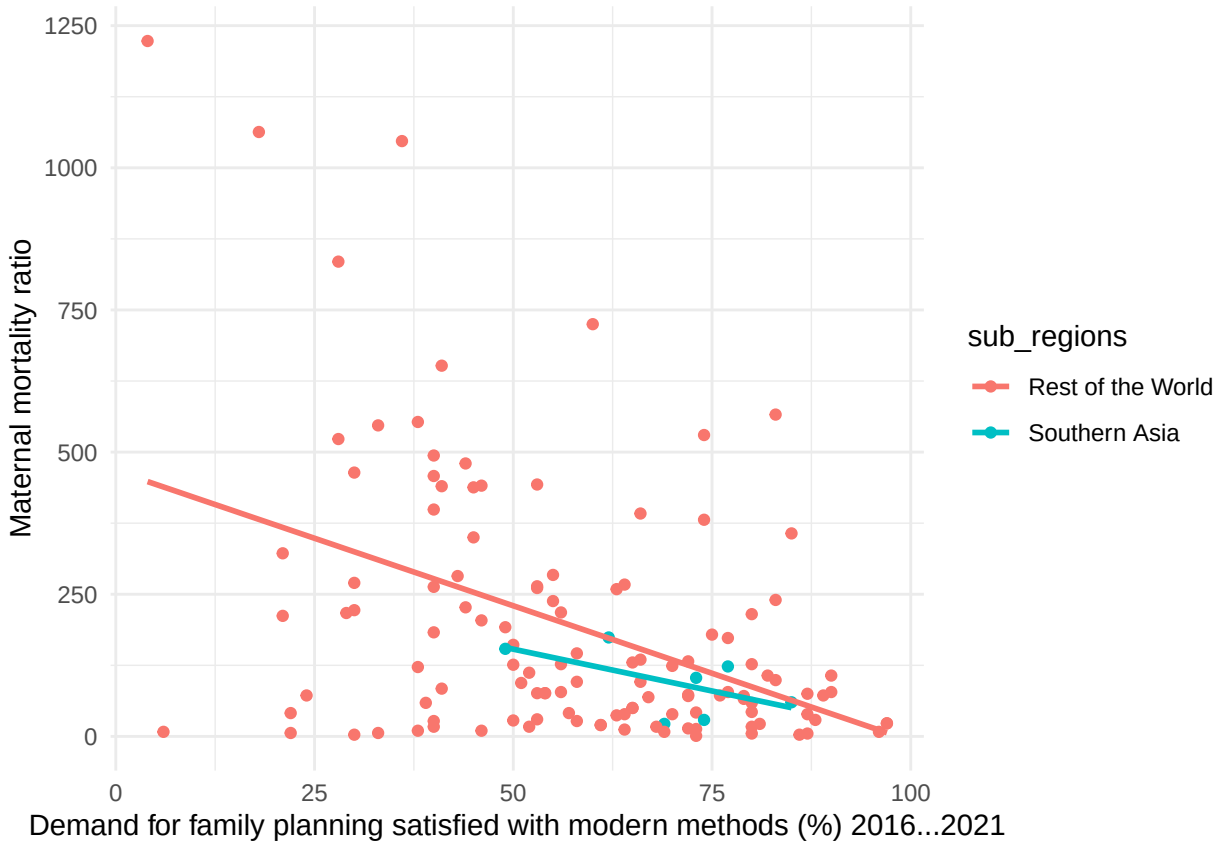
```
##
## Call:
## lm(formula = maternal_mortality_ratio ~ demand_family_planning,
##     data = rest_world)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -430.68 -123.49  -43.37   68.87  774.83
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              467.169     56.817   8.222 2.97e-13 ***
## demand_family_planning    -4.749      0.927  -5.123 1.19e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 210.3 on 118 degrees of freedom
## Multiple R-squared:  0.1819, Adjusted R-squared:  0.175
## F-statistic: 26.24 on 1 and 118 DF,  p-value: 1.188e-06
```

```r
cor(x=rest_world$demand_family_planning, y=rest_world$maternal_mortality_ratio)
```

```
## [1] -0.4265292
```

Based on external research (listed in poster), Afghanistan currently experiences many other circumstances which contribute to its unusually high maternal mortality ratio, such as high political instability (ie. involving the Taliban). Maldives also has unusually low maternal morality rate due to the extensive government effort to reduce maternal mortality. Maldives also has the highest GDP out of the other South Asian countries, at 3 times more than the second highest country (Bhutan). Thus, Afghanistan and Maldives are likely to be outliers.

```r
# ------------------------------------------------------------
# Create a dataset with outliers removed for South Asia
outliers_removed <- merged_data %>% filter(country_names != "Afghanistan" &
                                           country_names != "Maldives")

# Plot w/o outliers
outliers_removed %>% ggplot(aes(x=demand_family_planning, y=maternal_mortality_ratio,
                                color=sub_regions)) +
  geom_point() +
  labs(x="Demand for family planning satisfied with modern methods (%) 2016-2021",
       y="Maternal mortality ratio") +
  geom_smooth(method="lm", se=FALSE) +
  theme_minimal()
```




```
# Get correlation values on dataset with outliers removed
split_data <- split(outliers_removed, outliers_removed$sub_regions)
southern_asia = split_data[["Southern Asia"]]
rest_world = split_data[["Rest of the World"]]
```

```
# Southern Asia correlation + summary table:
cor(x=southern_asia$demand_family_planning, y=southern_asia$maternal_mortality_ratio)
```

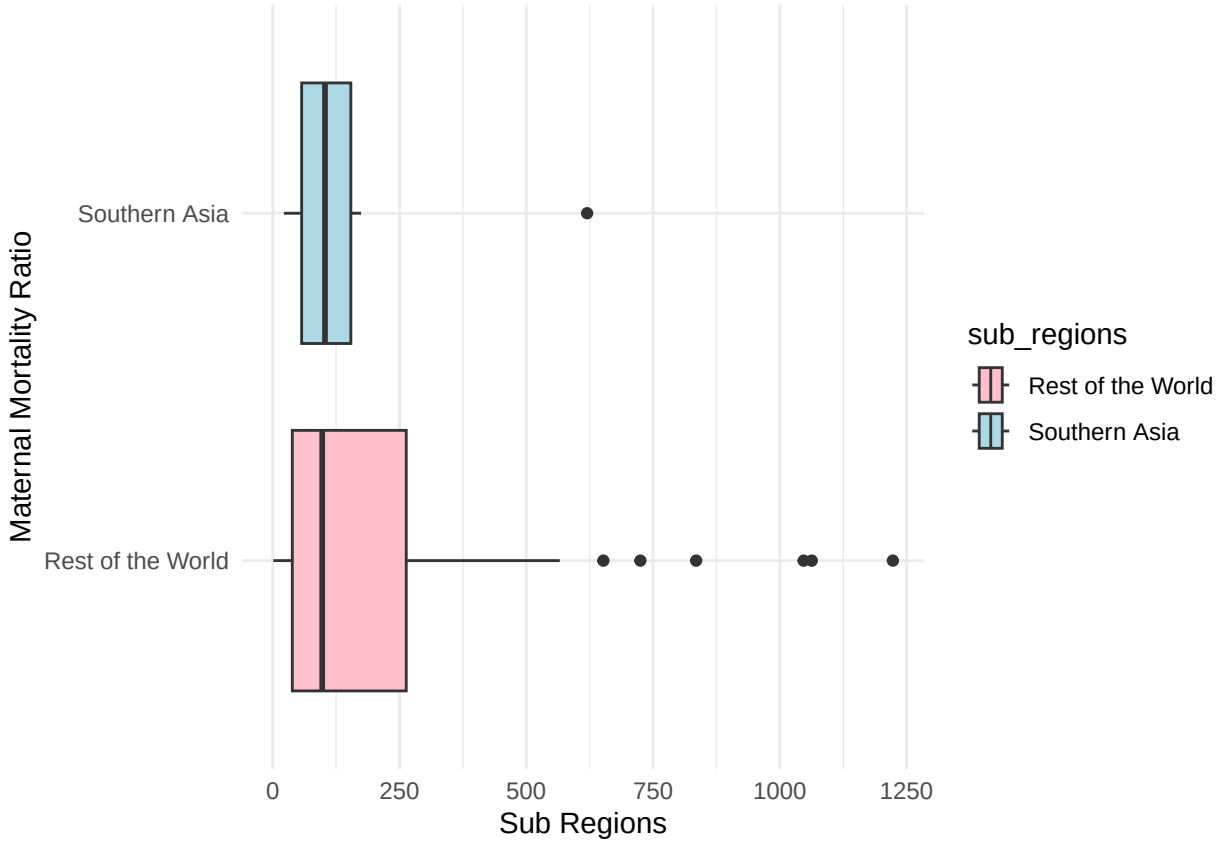
```
## [1] -0.5692745
```

```
model_southern_asia_no_outliers <- lm(maternal_mortality_ratio ~ demand_family_planning,
                                      data=southern_asia)

summary(model_southern_asia_no_outliers)
```

```
##
## Call:
## lm(formula = maternal_mortality_ratio ~ demand_family_planning,
##     data = southern_asia)
##
## Residuals:
##       1       2       3       4       5       6       7
##  49.006   9.534  17.243 -75.521  55.893  -2.339 -53.816
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               300.442    134.243   2.238   0.0754 .
## demand_family_planning     -2.941      1.899  -1.548   0.1822
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53.89 on 5 degrees of freedom
## Multiple R-squared:  0.3241, Adjusted R-squared:  0.1889
## F-statistic: 2.397 on 1 and 5 DF,  p-value: 0.1822
```

```r
# A side-by-side box plot of South Asian countries and the rest of the world based
# on maternal mortality ratio.
merged_data %>% ggplot(aes(x=maternal_mortality_ratio, y=sub_regions, fill=sub_regions)) +
  geom_boxplot() +
  labs(x="Sub Regions", y="Maternal Mortality Ratio") +
  scale_fill_manual(values = c("Southern Asia" = "lightblue", "Rest of the World" = "pink")) +
  theme_minimal()
```




```r
merged_data
```

```
## # A tibble: 129 x 5
##    country_names demand_family_planning maternal_mortality_r~1   `0` sub_regions
##    <chr>                          <dbl>                  <dbl> <dbl> <chr>
##  1 Afghanistan                       42                    620     0 Southern A~
##  2 Albania                            6                      8     0 Rest of th~
##  3 Algeria                           77                     78     0 Rest of th~
##  4 Angola                            30                    222     0 Rest of th~
##  5 Armenia                           40                     27     0 Rest of th~
##  6 Azerbaijan                        22                     41     0 Rest of th~
##  7 Bangladesh                        77                    123     0 Southern A~
##  8 Barbados                          70                     39     0 Rest of th~
```

```
##  9 Belarus                            73                     1     0 Rest of th~
## 10 Belize                             65                   130     0 Rest of th~
## # i 119 more rows
## # i abbreviated name: 1: maternal_mortality_ratio
```

Analysis of the Linear Regression Model

Fitted Regression Line for WITH outliers

Southern Asia: maternal_mortality_ratio = 416.70 - 4.30 demand_family_planning

Rest of the World: maternal_mortality_ratio = 467.169 - 4.749 demand_family_planning

Fitted Regression Line for Southern Asia WITHOUT outliers

Southern Asia: maternal_mortality_ratio = 300.442 - 2.941 demand_family_planning

As our research question involves comparing the relationship between family planning resources and maternal mortality ratio in Southern Asia and the rest of the world, performing the linear regression method is most suitable to make this comparison. The (univariate) linear regression method facilitates the comparison by providing separate coefficient values for each sub_region, which in fact allows us to make a direct assessment in the observed associations.

Coefficient r represents the strength of the correlation between the Maternal Mortality Ratio and Demand for family planning satisfied with modern methods. Based on the lm() regression model, including outliers, the correlation between the demand for family planning and maternal mortality ratio is approximately -0.430 for Southern Asia and -0.426 for the rest of the world. Excluding outliers, the correlation between the demand for family planning and maternal mortality ratio is approximately -0.569 and -0.426, respectively.

This implies that there appears to be a negative, moderate relationship between the variables. Likewise, beta_hat_1 represents how for every percent increase in Demand for family planning satisfied with modern methods, the Maternal Mortality Ratio decreases by a certain rate. Including outliers, on average, for each percent increase in Demand for family planning satisfied with modern methods, the Maternal Mortality Ratio in Southern Asia decreases by a rate of around 4.3 per 100,000 births and 4.75 per 100,000 births for the rest of the world. Removing the outliers, the maternal mortality ratio decreases by a rate of 2.94 and 4.75, respectively.

This satisfies our initial assumption that there would be lower maternal mortality for countries with a higher percentage of demand for family planning. As mentioned in the article published in thelancet.com, the external exposures contributing to the increase in maternal mortality ratio could be the result of insufficient healthcare, political, economical, cultural, and even lifestyle patterns.

In comparison to both linear regressions, it appears that Afghanistan was one of the 2 outliers amongst all other Southern Asian countries in the dataset. This could have been of other confounding variables and circumstances, such as illiteracy, unskilled staff, political instability due to the Taliban, and many more. According to the article published in sciencedirect.com, since health care resources have become inaccessible to most of the population, this resulted in around 4 million women and girls in dire need of humanitarian assistance as well as leading to additional 51,000 maternal deaths and 4.8 million unintended pregnancies by 2025.

Also, the Maldives Maternal Ratio is strikingly low compared with its Demand for family planning satisfied with modern methods as a Southern Asian country. According to an article published on linkedin.com, this is due to the Maldivian government having done extensive analysis on the issue and implementing multiple strategies to reduce their maternal mortality ratio. This includes increasing worker training and number of midwives, expanding obstetric resources, and increasing family planning education and contraceptive availability.

Analysis of the Side-by-Side Boxplot

To strengthen our analysis and have complementary insights of the linear regressions' data points, we chose to use a side-by-side boxplot to compare the maternal mortality ratio for the 2 sub_regions. This leads to a more visual overview and better communication in terms of identifying potential outliers, spread, mean, median, mode, and interquartile range (IQR).

Likewise, side-by-side boxplots help us visualize and interpret the differences there are in the distribution of the overall outcome across both sub_regions. In fact, this would also inform UNICEF whether or not there is a need for South Asian countries in particular to fund family planning resources to help reach the 3rd SDG Good Health and Well-being.

Although Afghanistan and Maldives appeared as the outliers in the linear regression model for Southern Asia, only Afghanistan is the outlier in the box plot. This is due to the fact that Maldives' Maternal Mortality ratio of 57 was rather fairly close to the maternal mortality ratio of the remaining Southern Asian countries as opposed to Afghanistan having a maternal mortality ratio of 620. Thus, this led for Maldives data point to be not as visible, so instead, it is within the range of the whiskers of the boxplot. The same reasoning to possible factors of Afghanistan being the outlier is mentioned above.

Conclusion

Finally, as Southern Asia has a higher correlation between the Maternal Mortality Ratio and Demand for family planning satisfaction, this means that Demand for family planning satisfaction has a more direct impact for Southern Asia.

However, according to the boxplot, the rest of the world's Maternal Mortality Ratio distribution is overall higher compared with Southern Asia. Thus, more funding should be used to increase Demand for family planning in the rest of the world since they have a greater need to reduce their Maternal Mortality Ratio.

Therefore, this should incentivize Non-Southern Asian countries to invest more in educational services regarding planned parenthood as well as health services and resources like a variety of modern contraceptives. However, notice that there are developed, industrialized countries that have little to no issue investing in family demands than developing countries due to their sophisticated economy, for instance the US and Myanmar respectively. Because of this, this imbalance is something UNICEF should take note of should they allocate fundings.

Potential Biases

Be that as it may, one limitation in this dataset would be that the sample size for Southern Asian countries consists of only 9 observations and only 8 with an outlier removed. By contrast, there consists of over 190 countries in the rest of the world. Since the sample size is relatively small, this could lead to imprecise estimates of the coefficient and limits the generalizability of our findings. The findings for Southern Asian countries may have been overshadowed by the much greater sample size of the rest of the world dataset, which leads to an unequal comparison.

Research Question 3

Are there specific groups of countries with similar progress towards the goal of Good Health and Well-being (SDG 3)? If so, how many groups are there and which groups do South Asian countries belong to?

First, we load in the Sustainable Development Report's SDG Index data.

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

```r
# 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"

# select relevant columns
sdg <-
  sdg %>%
  select(country_code_iso3, goal_3_score, goal_3_trend)  # remove first column

# preview data
sdg
```

```
## # A tibble: 206 x 3
##    country_code_iso3 goal_3_score goal_3_trend
##    <chr>                   <dbl> <chr>
##  1 FIN                      95.4 Score moderately improving, insufficient to a~
##  2 SWE                      96.9 Score moderately improving, insufficient to a~
##  3 DNK                      95.4 Score moderately improving, insufficient to a~
##  4 DEU                      93   Score moderately improving, insufficient to a~
##  5 AUT                      92.5 Score moderately improving, insufficient to a~
##  6 FRA                      93.2 Score moderately improving, insufficient to a~
##  7 NOR                      97.1 Score moderately improving, insufficient to a~
##  8 CZE                      90.2 Score moderately improving, insufficient to a~
##  9 POL                      85.2 Score moderately improving, insufficient to a~
## 10 EST                      89.5 Score moderately improving, insufficient to a~
## # i 196 more rows
```

Then, we load in the data for the country codes.

```r
# load country code data
country_codes <-
  read_csv("country_codes.csv") %>%
  select(-...1)  # remove first column

# rename columns
names(country_codes)[names(country_codes)=="ISO-alpha3 Code (M49)"] <-
  "country_code_iso3"
names(country_codes)[names(country_codes)=="Country or Area_en (M49)"] <-
  "country"
names(country_codes)[names(country_codes)=="Sub-region Name_en (M49)"] <-
```

```
  "subregion_name"

country_codes <-
  country_codes %>%
  select(country_code_iso3, country, subregion_name)

country_codes_south_asia <-
  country_codes %>%
  select(country_code_iso3, country, subregion_name) %>%
  filter(subregion_name == "Southern Asia")  # filter the data for South Asian countries

# preview data
country_codes
```

```
## # A tibble: 298 x 3
##    country_code_iso3 country                      subregion_name
##    <chr>             <chr>                        <chr>
##  1 DZA               Algeria                      Northern Africa
##  2 EGY               Egypt                        Northern Africa
##  3 LBY               Libya                        Northern Africa
##  4 MAR               Morocco                      Northern Africa
##  5 SDN               Sudan                        Northern Africa
##  6 TUN               Tunisia                      Northern Africa
##  7 ESH               Western Sahara               Northern Africa
##  8 IOT               British Indian Ocean Territory Sub-Saharan Africa
##  9 BDI               Burundi                      Sub-Saharan Africa
## 10 COM               Comoros                      Sub-Saharan Africa
## # i 288 more rows
```

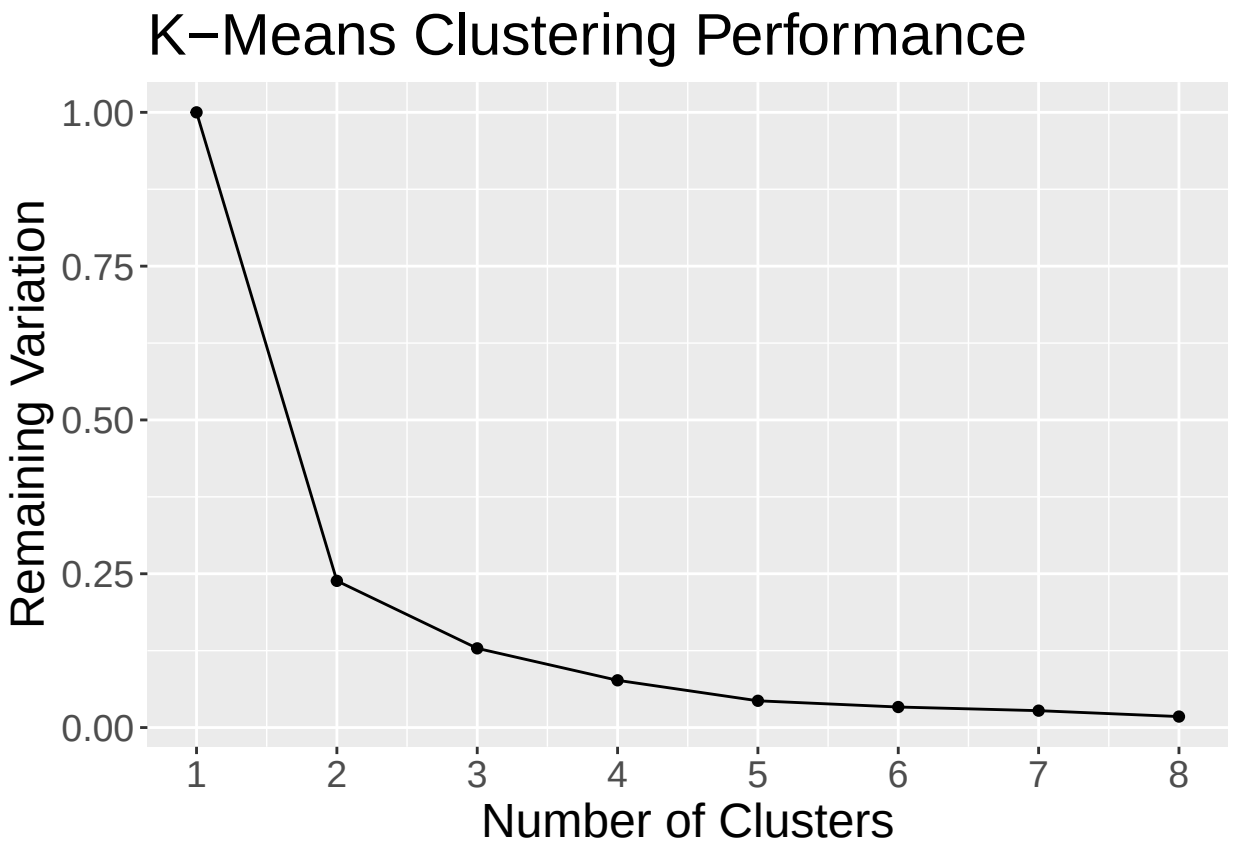
Next, we perform an inner join to combine the two tibbles.

```
# join tables
data <- inner_join(x=country_codes, y=sdg, by="country_code_iso3")

# filter countries without goal 3 score data
data <- data %>%
  filter(!is.na(goal_3_score))

# preview data
data
```

```
## # A tibble: 168 x 5
##    country_code_iso3 country  subregion_name     goal_3_score goal_3_trend
##    <chr>             <chr>    <chr>                     <dbl> <chr>
##  1 DZA               Algeria  Northern Africa            77.3 Score moderately ~
##  2 EGY               Egypt    Northern Africa            69.4 Score moderately ~
##  3 MAR               Morocco  Northern Africa            73.1 Score moderately ~
##  4 SDN               Sudan    Northern Africa            51.9 Score stagnating ~
##  5 TUN               Tunisia  Northern Africa            78.9 Score moderately ~
##  6 BDI               Burundi  Sub-Saharan Africa         48.5 Score stagnating ~
##  7 COM               Comoros  Sub-Saharan Africa         55.3 Score stagnating ~
##  8 DJI               Djibouti Sub-Saharan Africa         50.9 Score stagnating ~
##  9 ETH               Ethiopia Sub-Saharan Africa         46.5 Score moderately ~
## 10 KEN               Kenya    Sub-Saharan Africa         51   Score stagnating ~
## # i 158 more rows
```

Using the "elbow" approach, we find the optimal number of country clusters.

```
set.seed(130)
explained_ss <- rep(NA, 8)
for(k in 1:8){
  # run k-means on the data
  clustering2 <- kmeans(data$goal_3_score, k)
  explained_ss[k] <- clustering2$betweenss / clustering2$totss
}

# Plot evolution of metric as a function of k
ggplot() +
  aes(x=1:8, 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)) +
  scale_x_continuous(breaks=1:8) +
  scale_x_continuous(breaks=1:8)
```



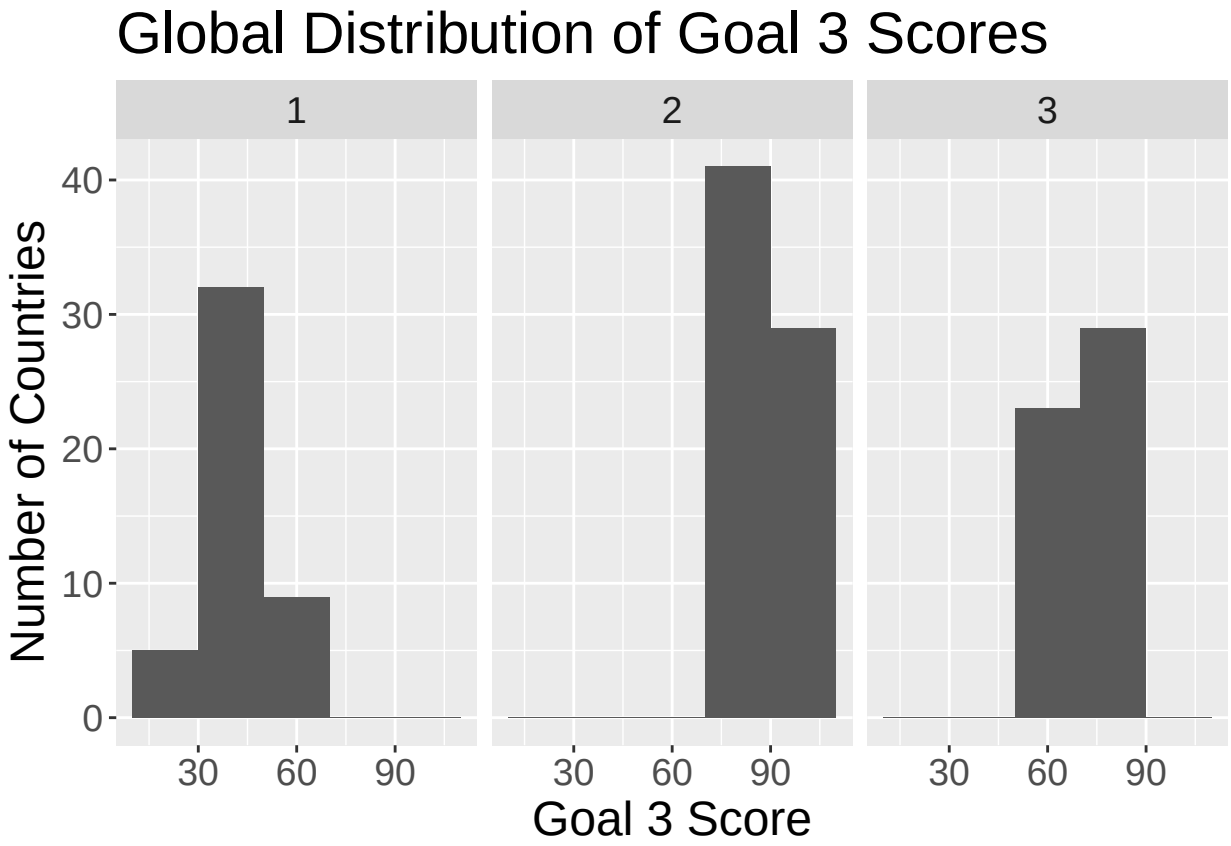
Since there is not a significant decrease in remaining variation as the number of clusters increases from 3 to 4, 3 is a good choice for the number of clusters. Adding another cluster will not improve the results much as compared to what we would get from random splits. We can conclude that there are three main groups of countries with similar progress towards the goal of Good Health and Well-being.

Then, we perform the clustering using 3 clusters.

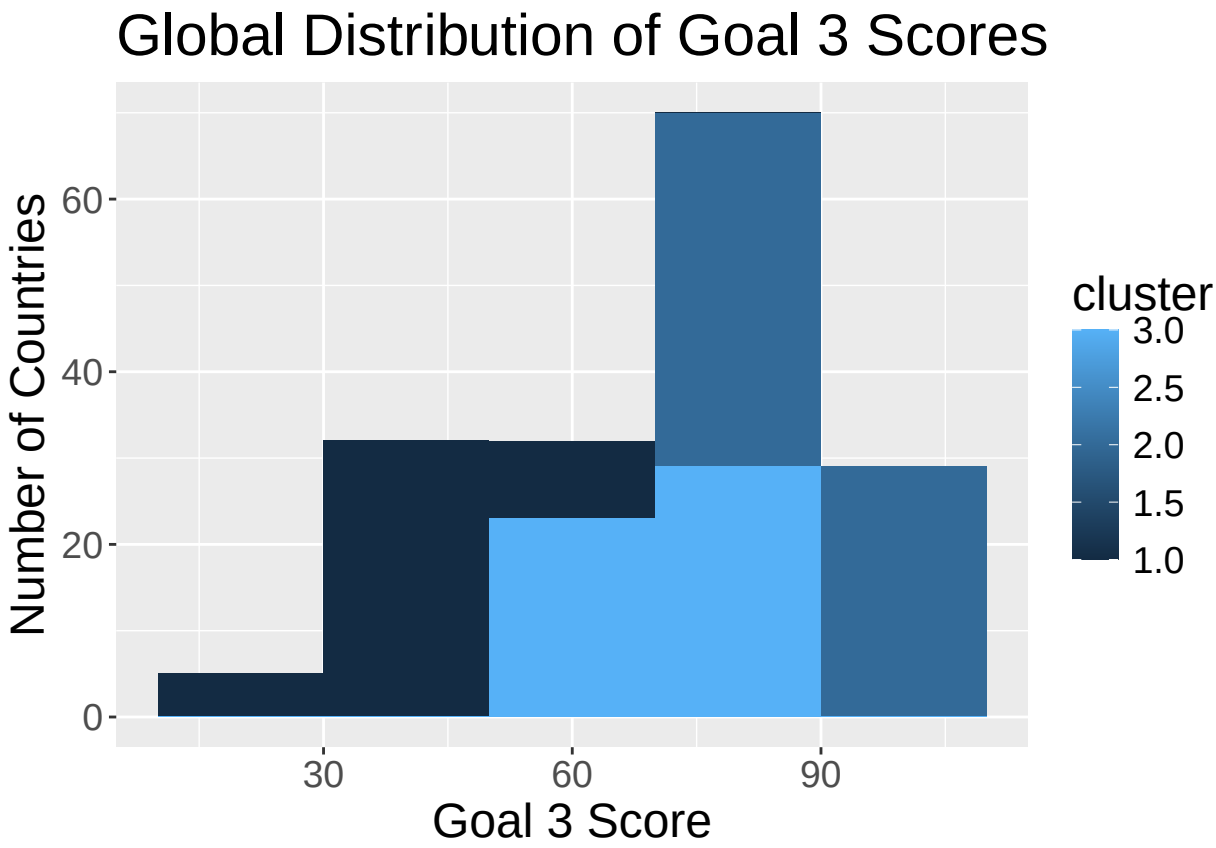
```r
set.seed(130)

# run k-means clustering with 3 clusters
clustering <- kmeans(data$goal_3_score, 3)
clustering
```

```
## K-means clustering with 3 clusters of sizes 46, 70, 52
##
## Cluster means:
##       [,1]
## 1 41.54565
## 2 87.95143
## 3 70.12500
##
## Clustering vector:
##   [1] 3 3 3 1 3 1 1 1 1 1 1 1 3 1 3 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 1 3 1 1
##  [38] 3 1 1 1 1 1 1 1 1 1 3 1 1 2 2 2 3 1 3 2 3 2 2 3 3 2 3 3 2 3 3 2 2 3 3 3 2
##  [75] 3 2 3 2 2 2 3 3 3 3 2 2 2 3 2 2 3 3 1 2 3 2 2 3 1 3 3 3 2 2 3 1 3 3 3 2 2
## [112] 3 3 3 2 3 2 3 2 2 2 3 2 2 1 2 2 2 2 2 3 2 2 2 3 2 2 2 2 2 2 2 2 2 2 3 2
## [149] 2 2 2 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 3 1
##
## Within cluster sum of squares by cluster:
## [1] 4280.794 2246.435 2307.298
##  (between_SS / total_SS =  87.1 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

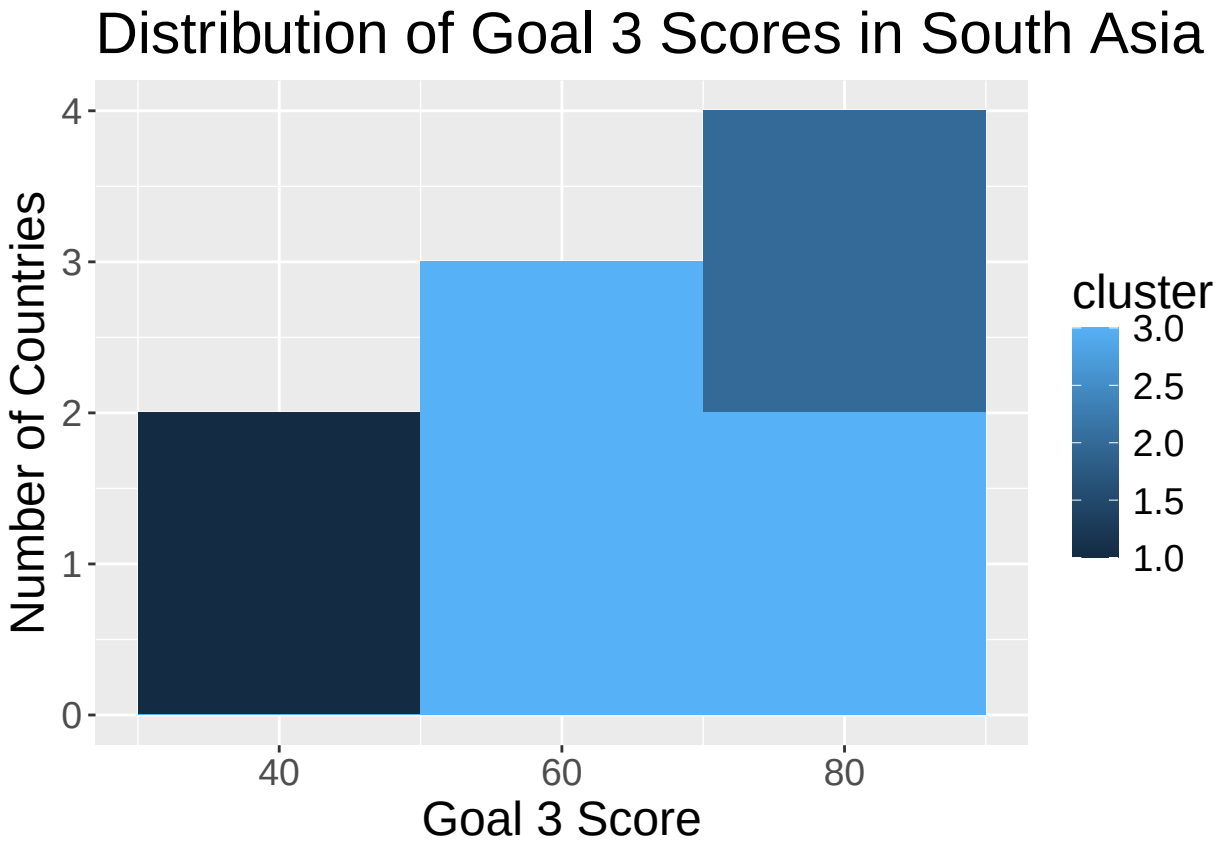
```r
# add clustering values to our original dataset
data <-
  data %>%
  mutate(cluster = clustering$cluster)

# filter the data for South Asian countries
south_asia_data <- data %>%
  filter(subregion_name == "Southern Asia")

# plot distribution of goal 3 scores (facet_wrap)
data %>% ggplot(aes(x=goal_3_score)) +
  geom_histogram(binwidth=20) +
  labs(x="Goal 3 Score",
       y="Number of Countries",
       title="Global Distribution of Goal 3 Scores") +
  theme(text=element_text(size=18)) +
  facet_wrap(~cluster)
```



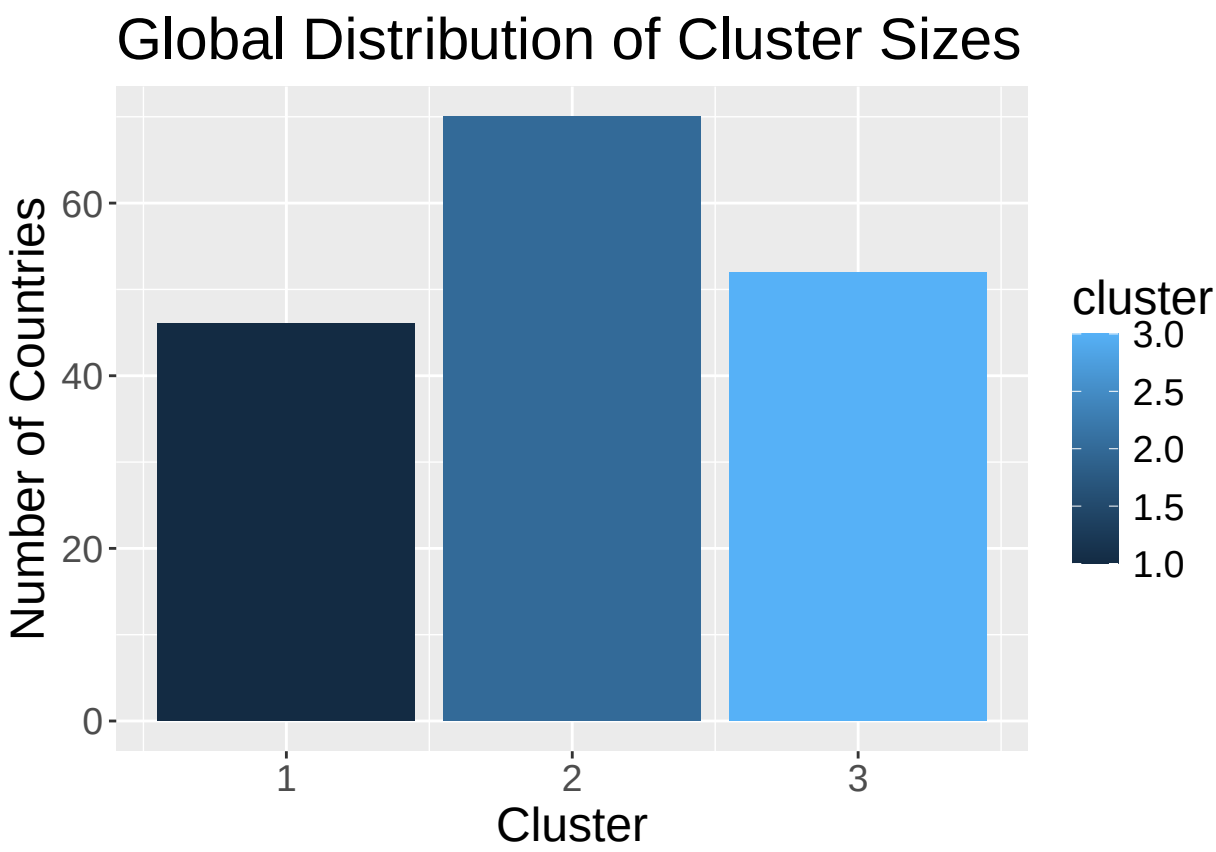

```
# plot distribution of goal 3 scores (color)
data %>% ggplot(aes(x=goal_3_score, group=cluster, fill=cluster)) +
  geom_histogram(binwidth=20) +
  labs(x="Goal 3 Score",
       y="Number of Countries",
       title="Global Distribution of Goal 3 Scores") +
  theme(text=element_text(size=18))
```



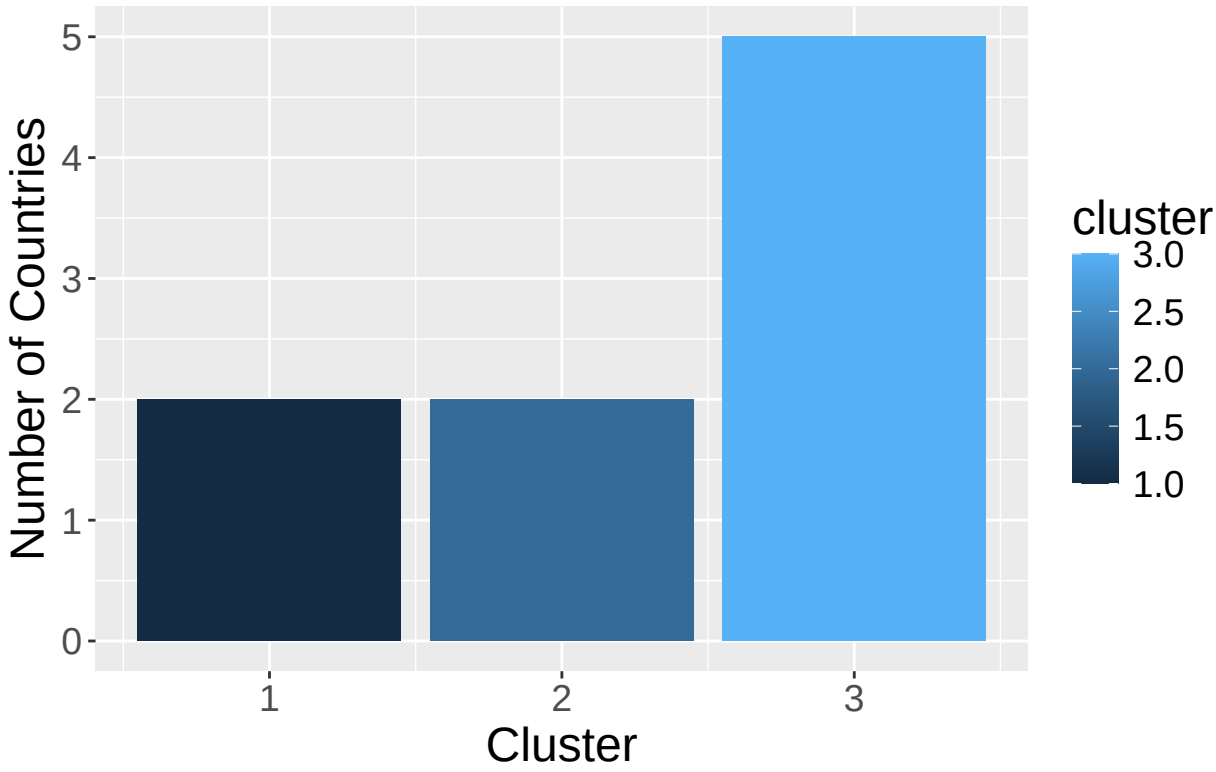

```
# plot distribution of goal 3 scores for South Asian countries (color)
south_asia_data %>% ggplot(aes(x=goal_3_score, group=cluster, fill=cluster)) +
  geom_histogram(binwidth=20) +
  labs(x="Goal 3 Score",
       y="Number of Countries",
       title="Distribution of Goal 3 Scores in South Asia") +
  theme(text=element_text(size=18))
```




```
# plot number of observations in each cluster
data %>%
  ggplot() +
  geom_bar(aes(x=cluster, group=cluster, fill=cluster)) +
  labs(x="Cluster",
       y="Number of Countries",
       title="Global Distribution of Cluster Sizes") +
  theme(text=element_text(size=18))
```




```
# plot number of observations in each cluster for South Asian countries
south_asia_data %>%
  ggplot() +
  geom_bar(aes(x=cluster, group=cluster, fill=cluster)) +
  labs(x="Cluster",
       y="Number of Countries",
       title="Distribution of Cluster Sizes in South Asia") +
  theme(text=element_text(size=18))
```



Based on the global distribution of Goal 3 scores, cluster 2 represents the countries with good progress, cluster 3 represents the countries with moderate progress, and cluster 1 represents the countries with slow progress. This implies that the countries in cluster 1 are the countries in most need of UNICEF interventions.

Globally, most countries belong to cluster 2, which shows that most countries are progressing well towards Goal 3. In South Asia, the situation is different - most countries belong to cluster 3, which means that most South Asian countries have moderate progress towards Goal 3. **It can be inferred that progress towards Goal 3 is generally slower in South Asia compared to the rest of the world.**

Lastly, we identify the clusters each South Asian country belongs to.

```
south_asia_data %>% select(country, goal_3_score, goal_3_trend, cluster)
```

```
## # A tibble: 9 x 4
##   country                    goal_3_score goal_3_trend                  cluster
##   <chr>                             <dbl> <chr>                          <int>
## 1 Afghanistan                        37.5 Score stagnating or increasin~       1
## 2 Bangladesh                         60.5 Score moderately improving, i~       3
## 3 Bhutan                             73.5 Score moderately improving, i~       3
## 4 India                              64.8 Score moderately improving, i~       3
## 5 Iran (Islamic Republic of)         80.6 Score moderately improving, i~       2
## 6 Maldives                           86.9 Score moderately improving, i~       2
## 7 Nepal                              59.6 Score moderately improving, i~       3
## 8 Pakistan                           45.2 Score stagnating or increasin~       1
## 9 Sri Lanka                          78.5 Score moderately improving, i~       3
```

By matching each country with their cluster, we see that the Maldives and Iran are progressing the best among South Asian countries. As seen from Goal 3 trend, the countries in cluster 1 (the "slow progress" cluster) also show stagnation in their scores. The problem of poor healthcare is particularly serious in Afghanistan

and Pakistan. **UNICEF could focus their efforts on countries in cluster 1 (Afghanistan, Pakistan) when designing programmes to promote good health and well-being in South Asia.**

Our results are also supported by other studies. According to data compiled from multiple sources by the World Bank as processed by Our World in Data (2023), for all South Asian countries apart from Iran and the Maldives, the number of medical doctors per 1,000 people is less than 1.5 in 2019. This implies that there is a lack of manpower in the healthcare system, which is a large barrier to accessible healthcare for the general population. Moreover, most South Asian countries are low in access to basic care, equality in healthcare services, and financial risk protection, which are the three main dimensions of universal health coverage. (Rahman et al., 2017) Another common issue is high disease rates but low health service coverage, which places a large burden on those who are unable to afford healthcare services. Therefore, we recommend that UNICEF consider allocating a larger budget to support poverty alleviation and healthcare inequality in South Asia.

Conclusion

To conclude, progress towards SDG3 is slower in South Asia compared to the rest of the world, especially in Afghanistan and Pakistan. UNICEF could allocate a larger budget to accelerate their contributions to SDG progress based on each group and help promote well-being and improve public health in South Asia.

Potential Biases

Lastly, we have to consider the potential limitations of using k-means clustering. The method of k-means clustering is based on the assumption that the size of each cluster is similar. (Robinson, 2015) If the size of the clusters are very uneven, then the k-means algorithm may not be able to find the clusters accurately.

References (APA Citation)

Research Question 1:

1. This article provides an in depth analysis on how U5MR is affected by various other factors. Madhav Kumar Bhusal and Shankar Prasad Khanal, Published online 2022 Dec 5, A Systematic Review of Factors Associated with Under-Five Child Mortality, Retrieved March 10, 2024 from A Systematic Review of Factors Associated with Under-Five Child Mortality - PMC https://www.ncbi.nlm.nih.gov /pmc/articles/PMC9744612/
2. This ChatGPT conversation helped in cleaning and formatting the file. https://chat.openai.com/shar e/8e2fb6aa-a2b1-482a-89d7-0b1b3144f2ca

Research Question 2:

1) We used the article published in lancet.com as a supporting argument that the outcome of the dataset satisfies our initial assumption that there would be lower maternal mortality for countries with a higher percentage of demand for family planning. The Lancet. (2019). The Lancet | The best science for better lives. https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(13)70059-7/fulltext
2) We used the articled published in sciencedirect.com to provide us with a more in-depth reasoning of why Afghanistan was one of the 2 outliers amongst all other Southern Asian countries in the dataset. Science Direct. (2022). ScienceDirect.com | Science, Health and Medical journals, Full Text Articles and books. Sciencedirect.com. https://www.sciencedirect.com/
3) The article posted on Linkedin provides us with a more in-depth reasoning of what external factors led Maldives to be the 2nd outlier aside from Afghanistan. How the Maldives Reduced Maternal Mortality by More Than 90 Percent — While the U.S. Rate Has Since Tripled. (n.d.). Www.linkedin.com. Retrieved April 5, 2024, from https://www.linkedin.com/pulse/how-maldives-reduced-maternal-mortality-more-than/

4) This is the conversation of how having a small dataset vs a large dataset influences the estimates of the coefficient values and generalization of the overall dataset. https://chat.openai.com/share/1cda8b6c-931a-4299-886c-84a6defeb5aa

5) This is the conversation of making our reasoning sound stronger on why we chose to use a side-by-side boxplot for maternal mortality ratio to provide us with complementary insights of the linear regression. https://chat.openai.com/share/c5f8976e-94e6-4a9f-8ba9-fdd82d3a8f67

Research Question 3:

1) We used data from this source to support our findings that progress towards SDG3 is slower in South Asia compared to the rest of the world.

Data compiled from multiple sources by World Bank – processed by Our World in Data (2023). "Medical doctors per 1,000 people" [dataset]. Data compiled from multiple sources by World Bank [original data]. https://ourworldindata.org/grapher/medical-doctors-per-1000-people-vs-gdp-per-capita

2) We referenced this article for information about the issues in the healthcare systems in South Asia.

Rahman, M. M., Karan, A., Rahman, M. S., Parsons, A., Abe, S. K., Bilano, V., Awan, R., Gilmour, S., & Shibuya, K. (2017). Progress Toward Universal Health Coverage: A Comparative Analysis in 5 South Asian Countries. JAMA internal medicine, 177(9), 1297–1305. https://doi.org/10.1001/jamainternmed.2017.3133

3) We referenced this article for the potential biases of k-means clustering.

Robinson, D. (2015, January 16). K-means clustering is not a free lunch. http://varianceexplained.org/r/kmeans-free-lunch/

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.