UNICEF Project Exploration - STA130 Winter 2024

Nikita Jain, Anish Pai, Anudari Jamsran

Final Project Overview: Identifying Opportunities to Accelerate Progress on Sustainable Development Goals (SDG)

Guiding Research Question

How do landlocked countries compare with small island nations in their advancement toward meeting the UN's Sustainable Development Goals (SDGs)?

Research Question 1:

Specific Research Question

Can we classify countries as having high or low economic growth based on different economic empowerment metrics (like education and time use), and which country type (landlocked vs. small island) has a lower error rate in this classification (higher accuracy)?

Data Wrangling and Cleaning

```
# Read in country codes data, select relevant columns for further analysis, drop any rows with missing
names <- read_csv("country_codes.csv")
```

```
## New names:
## Rows: 298 Columns: 125
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (99): Global Name_en (M49), Region Name_en (M49), Sub-region Name_en (M4... dbl
## (22): ...1, Global Code (M49), Region Code (M49), Intermediate Region Co... lgl
## (4): Sub-region Code (M49), Least Developed Countries (LDC) (M49), Land...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
names_ll <- names %>%
  select("ISO-alpha3 Code (M49)", "Country or Area_en (M49)", "Developed / Developing Countries (M49)",

# Similar to above, but for Small Island Developing States (SIDS) instead of landlocked countries.
names_si <- names %>%
  select("ISO-alpha3 Code (M49)", "Country or Area_en (M49)", "Developed / Developing Countries (M49)",
```

```
# Load in SDG goal scores, select relevant columns, drop rows with missing values, and rename the count
sdg <- read_csv("sdr_fd5e4b5a.csv") %>% select("Goal 8 Score", "country_label", "Country Code ISO3") %>%
```

```
## New names:
## Rows: 206 Columns: 59
## -- Column specification
```

```
## ------------------------------------------------------- Delimiter: "," chr
## (36): Goal 1 Dash, Goal 1 Trend, Goal 2 Dash, Goal 2 Trend, Goal 3 Dash,... dbl
## (23): ...1, Goal 1 Score, Goal 2 Score, Goal 3 Score, Goal 4 Score, Goal...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
# Perform inner joins to match country codes from the names_ll and names_si datasets with their corresp
namell_sdg <- inner_join(x=names_ll, y=sdg, by="con_codes")
namesi_sdg <- inner_join(x=names_si, y=sdg, by="con_codes")
```

```
# Read in country indicators, select relevant columns for gender-related economic empowerment, drop row
indicators <- read_csv("country_indicators.csv") %>% select("sowc_women-s-economic-empowerment__educati
```

```
## New names:
## Rows: 218 Columns: 1332
## -- Column specification
## ------------------------------------------------------- Delimiter: "," chr
## (8): iso3, hdr_hdicode, hdr_region, wbi_income_group, wbi_lending_cat... dbl
## (1324): ...1, sowc_demographics__population-thousands-2021_total, sowc_d...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
# Merge the country indicators with the SDG scores for both landlocked and small island countries, excl
data_land_locked <- inner_join(x=namell_sdg, y=indicators, by="con_codes") %>% select(-`Developed / Dev
data_small_island <- inner_join(x=namesi_sdg, y=indicators, by="con_codes") %>% select(-`Developed / De
```

```
# Calculate the mean Goal 8 Score for both small island and landlocked country groups. This mean will b
smallmean = mean(data_small_island$`Goal 8 Score`, na.rm = TRUE)
landmean = mean(data_land_locked$`Goal 8 Score`, na.rm = TRUE)

# For landlocked countries, classify economic growth, calculate average educational attainment, labor f
data_land_locked <- data_land_locked %>%
  mutate(EconomicGrowthLabel = ifelse(`Goal 8 Score` >= ((smallmean + landmean) / 2), 'good', 'bad')) %
  mutate(AvgEducationalAttainment = (`sowc_women-s-economic-empowerment__educational-attainment-2008-20
  mutate(AvgLaborForceParticipation = (`sowc_women-s-economic-empowerment__labour-force-participation-ra
                                       `sowc_women-s-economic-empowerment__labour-force-participation-ra
  mutate(AvgUnemploymentRate = (`sowc_women-s-economic-empowerment__unemployment-rate-2010-2020-r_femal
                                       `sowc_women-s-economic-empowerment__unemployment-rate-2010-2020-r

data_small_island <- data_small_island %>%
  mutate(EconomicGrowthLabel = ifelse(`Goal 8 Score` >= ((smallmean + landmean) / 2), 'good', 'bad')) %
                                      `sowc_women-s-economic-empowerment__educational-attainment-2008-20
  mutate(AvgLaborForceParticipation = (`sowc_women-s-economic-empowerment__labour-force-participation-ra
                                       `sowc_women-s-economic-empowerment__labour-force-participation-ra
  mutate(AvgUnemploymentRate = (`sowc_women-s-economic-empowerment__unemployment-rate-2010-2020-r_femal
                                       `sowc_women-s-economic-empowerment__unemployment-rate-2010-2020-r
```

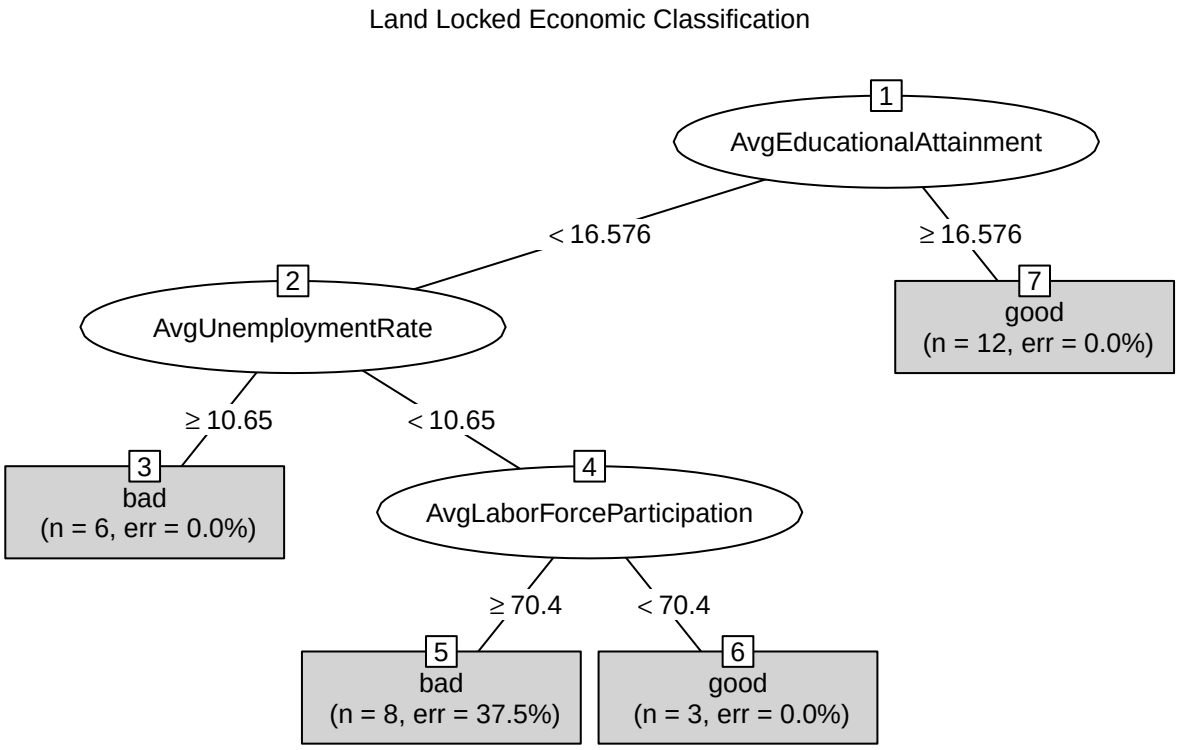
Tree Creation

```
# Build a classification tree model (tree_model1) for landlocked countries. The model predicts the Econ
tree_model1 <- rpart(EconomicGrowthLabel ~
                       AvgEducationalAttainment + AvgUnemploymentRate
                     + AvgLaborForceParticipation,
```

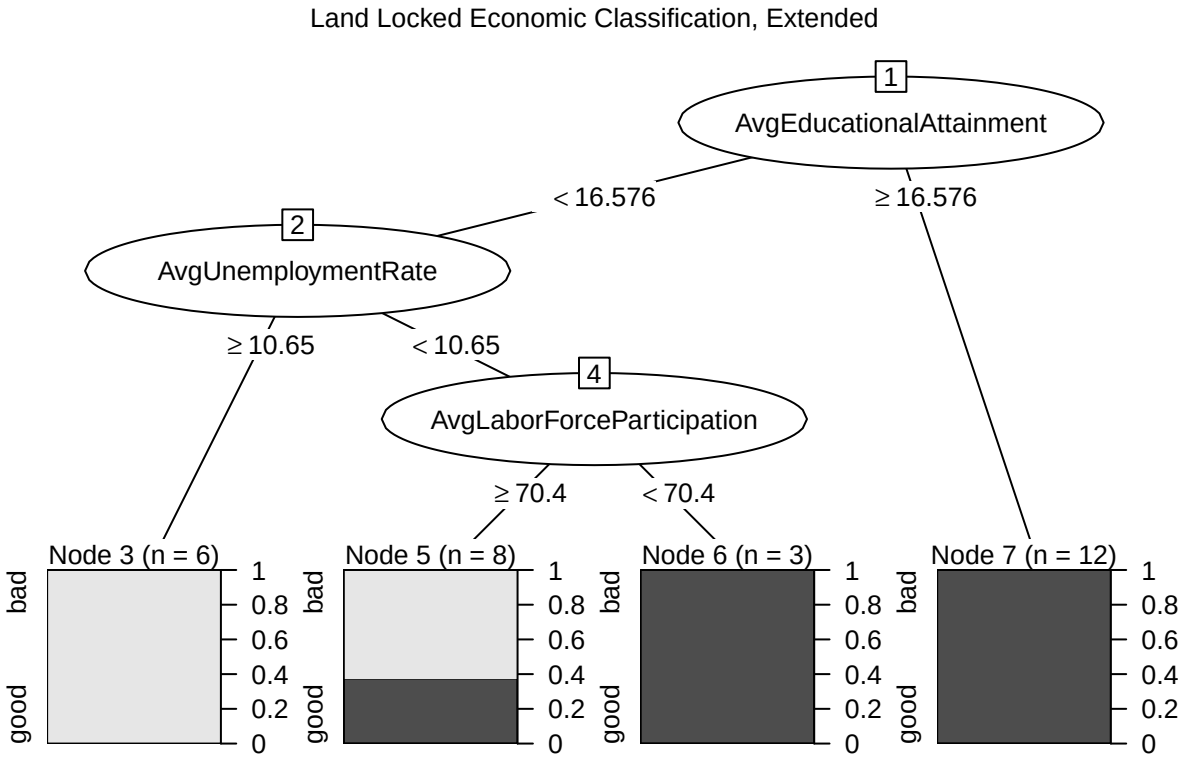
```
                    data = data_land_locked,
                    method = "class",
                    control = rpart.control(cp = 0.1,
                                            minsplit = 5,
                                            minbucket = 1,
                                            maxdepth = 30))

# Similarly, build a classification tree model (tree_model2) for small island countries using the same

tree_model2 <- rpart(EconomicGrowthLabel ~
                       AvgEducationalAttainment +
                       AvgUnemploymentRate + AvgLaborForceParticipation,
                     data = data_small_island,
                     method = "class",
                     control = rpart.control(cp = 0.1,
                                             minsplit = 5,
                                             minbucket = 1,
                                             maxdepth = 30))
```

Visualizations

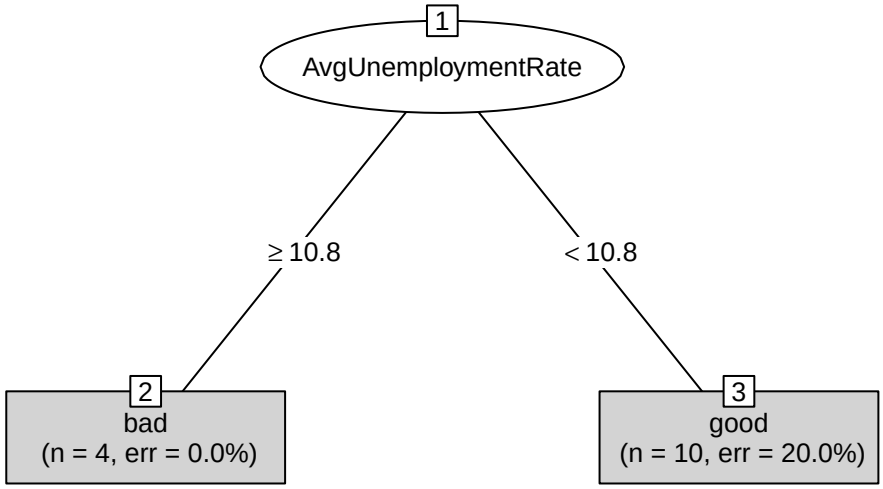
```
# Visualize the classification tree (tree_model1) for landlocked countries. Two types of plots are gene
plot(as.party(tree_model1), type = "simple", gp=gpar(cex=0.8), main = "Land Locked Economic Classificat
```



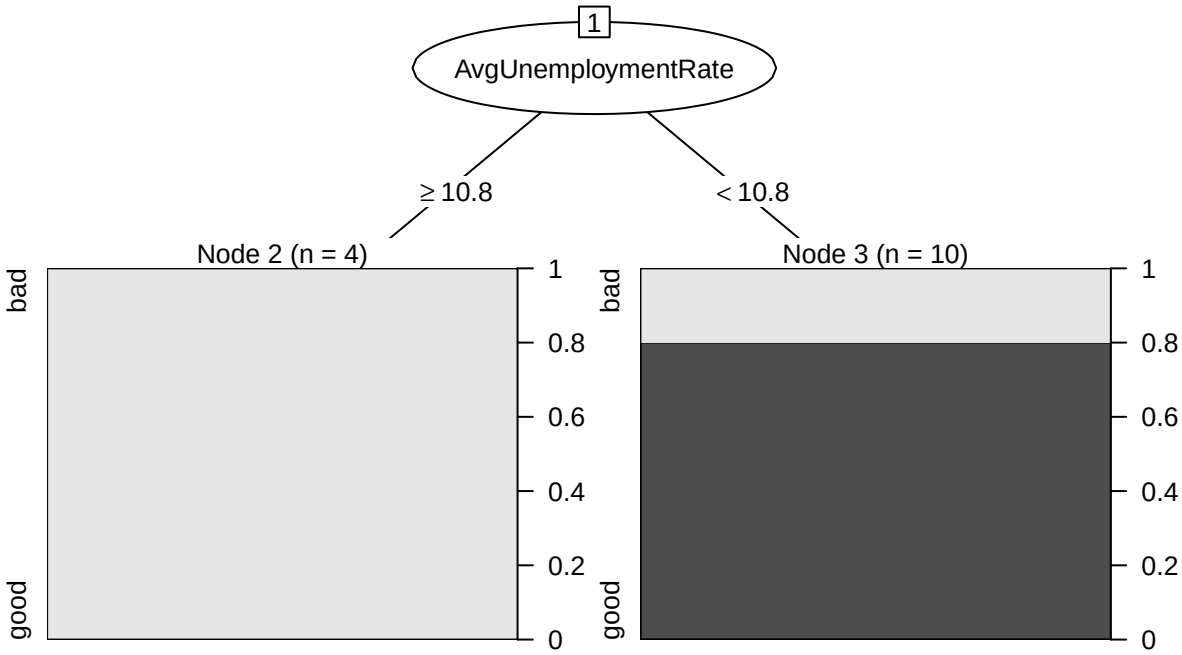

```
plot(as.party(tree_model1), type = "extended", gp=gpar(cex=0.8), main = "Land Locked Economic Classifica
```




```
# Visualize the classification tree (tree_model2) for small island countries in the same manner as for
plot(as.party(tree_model2), type = "simple", gp=gpar(cex=0.8), main = "Small Island Economic Classifica
```




```
plot(as.party(tree_model2), type = "extended", gp=gpar(cex=0.8), main = "Small Island Economic Classifi
```



Research Question 2

```
# load in country indicators
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
```

```
## New names:
## Rows: 218 Columns: 1332
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (8): iso3, hdr_hdicode, hdr_region, wbi_income_group, wbi_lending_cat... dbl
## (1324): ...1, sowc_demographics__population-thousands-2021_total, sowc_d...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
# preview data
country_indicators
```

```
## # A tibble: 218 x 1,331
##    country_code_iso3 sowc_demographics__population-thou~1 sowc_demographics__p~2
##    <chr>                                           <dbl>                  <dbl>
##  1 AFG                                            40099.                 20298.
##  2 ALB                                             2855.                   574.
##  3 DZA                                            44178.                 15526.
##  4 AND                                               79.0                   12.8
##  5 AGO                                            34504.                 17833.
##  6 AIA                                               15.8                    3.29
```

```
##  7 ATG                                                  93.2                  21.3
##  8 ARG                                                45277.                12669.
##  9 ARM                                                 2791.                  669.
## 10 AUS                                                25921.                 5667.
## # i 208 more rows
## # i abbreviated names: 1: `sowc_demographics__population-thousands-2021_total`,
## #   2: `sowc_demographics__population-thousands-2021_under-18`
## # i 1,328 more variables:
## #   `sowc_demographics__population-thousands-2021_under-5` <dbl>,
## #   `sowc_demographics__annual-population-growth-rate_2000-2020` <dbl>,
## #   `sowc_demographics__annual-population-growth-rate_2020-2030-a` <dbl>, ...
```

We see in addition to the country codes, which we have called `country_code_iso3`, there is also a whole host of additional information on each country. A list of codes from above is printed out below.

```
country_indicators$country_code_iso3
```

```
##   [1] "AFG" "ALB" "DZA" "AND" "AGO" "AIA" "ATG" "ARG" "ARM" "AUS" "AUT" "AZE"
##  [13] "BHS" "BHR" "BGD" "BRB" "BLR" "BEL" "BLZ" "BEN" "BTN" "BOL" "BIH" "BWA"
##  [25] "BRA" "VGB" "BRN" "BGR" "BFA" "BDI" "CPV" "KHM" "CMR" "CAN" "CAF" "TCD"
##  [37] "CHL" "CHN" "COL" "COM" "COG" "COK" "CRI" "CIV" "HRV" "CUB" "CYP" "CZE"
##  [49] "PRK" "COD" "DNK" "DJI" "DMA" "DOM" "ECU" "EGY" "SLV" "GNQ" "ERI" "EST"
##  [61] "SWZ" "ETH" "FJI" "FIN" "FRA" "GAB" "GMB" "GEO" "DEU" "GHA" "GRC" "GRD"
##  [73] "GTM" "GIN" "GNB" "GUY" "HTI" "VAT" "HND" "HUN" "ISL" "IND" "IDN" "IRN"
##  [85] "IRQ" "IRL" "ISR" "ITA" "JAM" "JPN" "JOR" "KAZ" "KEN" "KIR" "KWT" "KGZ"
##  [97] "LAO" "LVA" "LBN" "LSO" "LBR" "LBY" "LIE" "LTU" "LUX" "MDG" "MWI" "MYS"
## [109] "MDV" "MLI" "MLT" "MHL" "MRT" "MUS" "MEX" "FSM" "MCO" "MNG" "MNE" "MSR"
## [121] "MAR" "MOZ" "MMR" "NAM" "NRU" "NPL" "NLD" "NZL" "NIC" "NER" "NGA" "NIU"
## [133] "MKD" "NOR" "OMN" "PAK" "PLW" "PAN" "PNG" "PRY" "PER" "PHL" "POL" "PRT"
## [145] "QAT" "KOR" "MDA" "ROU" "RUS" "RWA" "KNA" "LCA" "VCT" "WSM" "SMR" "STP"
## [157] "SAU" "SEN" "SRB" "SYC" "SLE" "SGP" "SVK" "SVN" "SLB" "SOM" "ZAF" "SSD"
## [169] "ESP" "LKA" "PSE" "SDN" "SUR" "SWE" "CHE" "SYR" "TJK" "THA" "TLS" "TGO"
## [181] "TKL" "TON" "TTO" "TUN" "TKM" "TCA" "TUV" "UGA" "UKR" "ARE" "GBR" "TZA"
## [193] "USA" "URY" "UZB" "VUT" "VEN" "VNM" "YEM" "ZMB" "ZWE" "TUR" "ABW" "ASM"
## [205] "BMU" "CUW" "CYM" "FRO" "GIB" "GRL" "GUM" "IMN" "MNP" "NCL" "PRI" "PYF"
## [217] "SXM" "XKX"
```

Next let's take a look at the Sustainable Development Report's SDG Index data.

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

```
## New names:
## Rows: 206 Columns: 59
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (36): Goal 1 Dash, Goal 1 Trend, Goal 2 Dash, Goal 2 Trend, Goal 3 Dash,... dbl
## (23): ...1, Goal 1 Score, Goal 2 Score, Goal 3 Score, Goal 4 Score, Goal...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
# 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"

# preview data
sdg
```

```
## # A tibble: 206 x 58
##    goal_1_status     goal_1_trend       goal_2_status goal_2_trend goal_3_status
##    <chr>             <chr>              <chr>         <chr>        <chr>
##  1 SDG achieved      On track or maint~ Major challe~ Score stagn~ Challenges r~
##  2 SDG achieved      On track or maint~ Major challe~ Score stagn~ Challenges r~
##  3 SDG achieved      Score moderately ~ Significant ~ Score stagn~ Challenges r~
##  4 Challenges remain Decreasing         Significant ~ Score stagn~ Significant ~
##  5 SDG achieved      Score moderately ~ Significant ~ Score stagn~ Challenges r~
##  6 SDG achieved      Score moderately ~ Significant ~ Score stagn~ Significant ~
##  7 SDG achieved      Score stagnating ~ Major challe~ Score stagn~ SDG achieved
##  8 SDG achieved      On track or maint~ Major challe~ Score stagn~ Significant ~
##  9 SDG achieved      On track or maint~ Major challe~ Score stagn~ Significant ~
## 10 Challenges remain Score moderately ~ Major challe~ Score stagn~ Significant ~
## # i 196 more rows
## # i 53 more variables: goal_3_trend <chr>, goal_4_status <chr>,
## #   goal_4_trend <chr>, goal_5_status <chr>, goal_5_trend <chr>,
## #   goal_6_status <chr>, goal_6_trend <chr>, goal_7_status <chr>,
## #   goal_7_trend <chr>, goal_8_status <chr>, goal_8_trend <chr>,
## #   goal_9_status <chr>, goal_9_trend <chr>, goal_10_status <chr>,
## #   goal_10_trend <chr>, goal_11_status <chr>, goal_11_trend <chr>, ...
```

Joining the two datas together

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

# preview data
data
```

```
## # A tibble: 193 x 1,388
##    country_code_iso3 sowc_demographics__population-thou~1 sowc_demographics__p~2
##    <chr>                                           <dbl>                  <dbl>
##  1 AFG                                           40099.                 20298.
##  2 ALB                                            2855.                   574.
##  3 DZA                                           44178.                 15526.
##  4 AND                                              79.0                   12.8
```

```
##  5 AGO                                                34504.                 17833.
##  6 ATG                                                   93.2                   21.3
##  7 ARG                                                45277.                 12669.
##  8 ARM                                                 2791.                   669.
##  9 AUS                                                25921.                  5667.
## 10 AUT                                                 8922.                  1542.
## # i 183 more rows
## # i abbreviated names: 1: `sowc_demographics__population-thousands-2021_total`,
## #   2: `sowc_demographics__population-thousands-2021_under-18`
## # i 1,385 more variables:
## #   `sowc_demographics__population-thousands-2021_under-5` <dbl>,
## #   `sowc_demographics__annual-population-growth-rate_2000-2020` <dbl>,
## #   `sowc_demographics__annual-population-growth-rate_2020-2030-a` <dbl>, ...
```

Research Question 2: SDG 6

Goal 6: Clean water and sanitation Are people from landlocked countries using more clean waters compared to small island country people?

```
# From country_codes file select relevant columns for SDG 6
country_codes <- read_csv("country_codes.csv") %>%
  select("ISO-alpha3 Code (M49)",
         "Land Locked Developing Countries (LLDC) (M49)",
         "Small Island Developing States (SIDS) (M49)")
```

```
## New names:
## Rows: 298 Columns: 125
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (99): Global Name_en (M49), Region Name_en (M49), Sub-region Name_en (M4... dbl
## (22): ...1, Global Code (M49), Region Code (M49), Intermediate Region Co... lgl
## (4): Sub-region Code (M49), Least Developed Countries (LDC) (M49), Land...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
filtered_countries <- country_codes %>%
  filter(`Land Locked Developing Countries (LLDC) (M49)` == "TRUE" |
           `Small Island Developing States (SIDS) (M49)` == "TRUE")

print(filtered_countries)
```

```
## # A tibble: 85 x 3
##    `ISO-alpha3 Code (M49)` Land Locked Developing Count~1 Small Island Develop~2
##    <chr>                   <lgl>                          <lgl>
##  1 BDI                     TRUE                           NA
##  2 COM                     NA                             TRUE
##  3 ETH                     TRUE                           NA
##  4 MWI                     TRUE                           NA
##  5 MUS                     NA                             TRUE
##  6 RWA                     TRUE                           NA
##  7 SYC                     NA                             TRUE
##  8 SSD                     TRUE                           NA
##  9 UGA                     TRUE                           NA
## 10 ZMB                     TRUE                           NA
## # i 75 more rows
## # i abbreviated names: 1: `Land Locked Developing Countries (LLDC) (M49)`,
## #   2: `Small Island Developing States (SIDS) (M49)`
```

```
# For SDG 6 select needed data
data1 <- read_csv("country_indicators.csv") %>%
  select(-...1) %>%
  select(iso3, "sowc_wash__households-2020_at-least-basic-drinking-water-services_total",
         "sowc_wash__households-2020_at-least-basic-drinking-water-services_urban",
         "sowc_wash__households-2020_at-least-basic-drinking-water-services_rural") %>%
  rename(country_code_iso3 = iso3)
```

```
## New names:
## Rows: 218 Columns: 1332
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
```

```
## (8): iso3, hdr_hdicode, hdr_region, wbi_income_group, wbi_lending_cat... dbl
## (1324): ...1, sowc_demographics__population-thousands-2021_total, sowc_d...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
# Inner join data1 with filtered_countries
data2 <- inner_join(data1, filtered_countries,
                    by = c("country_code_iso3" = "ISO-alpha3 Code (M49)"))

print(data2)
```

```
## # A tibble: 83 x 6
##    country_code_iso3 sowc_wash__households-2020_at-leas~1 sowc_wash__household~2
##    <chr>                                           <dbl>                  <dbl>
##  1 AFG                                              75.1                  100
##  2 AIA                                              NA                     NA
##  3 ATG                                              NA                     NA
##  4 ARM                                             100.                   100.
##  5 AZE                                              96.0                  100
##  6 BHS                                              NA                     NA
##  7 BRB                                              98.5                   NA
##  8 BLZ                                              98.4                   98.9
##  9 BTN                                              97.3                   98.1
## 10 BOL                                              93.4                   99.1
## # i 73 more rows
## # i abbreviated names:
## #   1: `sowc_wash__households-2020_at-least-basic-drinking-water-services_total`,
## #   2: `sowc_wash__households-2020_at-least-basic-drinking-water-services_urban`
## # i 3 more variables:
## #   `sowc_wash__households-2020_at-least-basic-drinking-water-services_rural` <dbl>,
## #   `Land Locked Developing Countries (LLDC) (M49)` <lgl>, ...
```

```
# Select SDG goal 6 score from sdg dataset
sdg_goal_6 <- sdg %>%
  select(country_code_iso3, goal_6_score)

data3 <- left_join(data2, sdg_goal_6, by = "country_code_iso3")
names(data3)[names(data3) == "sowc_wash__households-2020_at-least-basic-drinking-water-services_total"]
  "clean_water_usage_total"
names(data3)[names(data3) == "sowc_wash__households-2020_at-least-basic-drinking-water-services_urban"]
  "clean_water_usage_urban"
names(data3)[names(data3) == "sowc_wash__households-2020_at-least-basic-drinking-water-services_rural"]
  "clean_water_usage_rural"
print(data3)
```

```
## # A tibble: 83 x 7
##    country_code_iso3 clean_water_usage_total clean_water_usage_urban
##    <chr>                               <dbl>                   <dbl>
##  1 AFG                                  75.1                   100
##  2 AIA                                  NA                      NA
##  3 ATG                                  NA                      NA
##  4 ARM                                 100.                    100.
##  5 AZE                                  96.0                   100
##  6 BHS                                  NA                      NA
##  7 BRB                                  98.5                    NA
```

```
##  8 BLZ                                  98.4                   98.9
##  9 BTN                                  97.3                   98.1
## 10 BOL                                  93.4                   99.1
## # i 73 more rows
## # i 4 more variables: clean_water_usage_rural <dbl>,
## #   `Land Locked Developing Countries (LLDC) (M49)` <lgl>,
## #   `Small Island Developing States (SIDS) (M49)` <lgl>, goal_6_score <dbl>
```

```r
seed <- 230
set.seed(seed)

n_trials <- 10000
n_sample <- 100

# Simulate clean water usage
clean_water_usage_simulations <- numeric(n_trials)

for (i in 1:n_trials) {
  clean_water_usage <- runif(n_sample, min = 0, max = 100)

  mean_clean_water_usage <- mean(clean_water_usage)

  clean_water_usage_simulations[i] <- mean_clean_water_usage
}
```

Hypothesis testing

```r
data3 <- data3 %>%
  mutate(Group = case_when(
    `Land Locked Developing Countries (LLDC) (M49)` == TRUE ~ "Landlocked",
    `Small Island Developing States (SIDS) (M49)` == TRUE ~ "Small Island",
    TRUE ~ "Other"
  ))

# Filter data to include only landlocked and small island countries
data3 <- data3 %>%
  filter(Group %in% c("Landlocked", "Small Island"))
data3 <- data3 %>%
  filter(!is.na(clean_water_usage_total), is.finite(clean_water_usage_total))

# Perform two-sample t-test
t_test_water <- t.test(clean_water_usage_total ~ Group, data = data3)

print(t_test_water)
```

```
##
##  Welch Two Sample t-test
##
## data:  clean_water_usage_total by Group
## t = -3.3355, df = 54.489, p-value = 0.001538
## alternative hypothesis: true difference in means between group Landlocked and group Small Island is n
## 95 percent confidence interval:
##  -22.950244  -5.720486
## sample estimates:
##   mean in group Landlocked mean in group Small Island
```

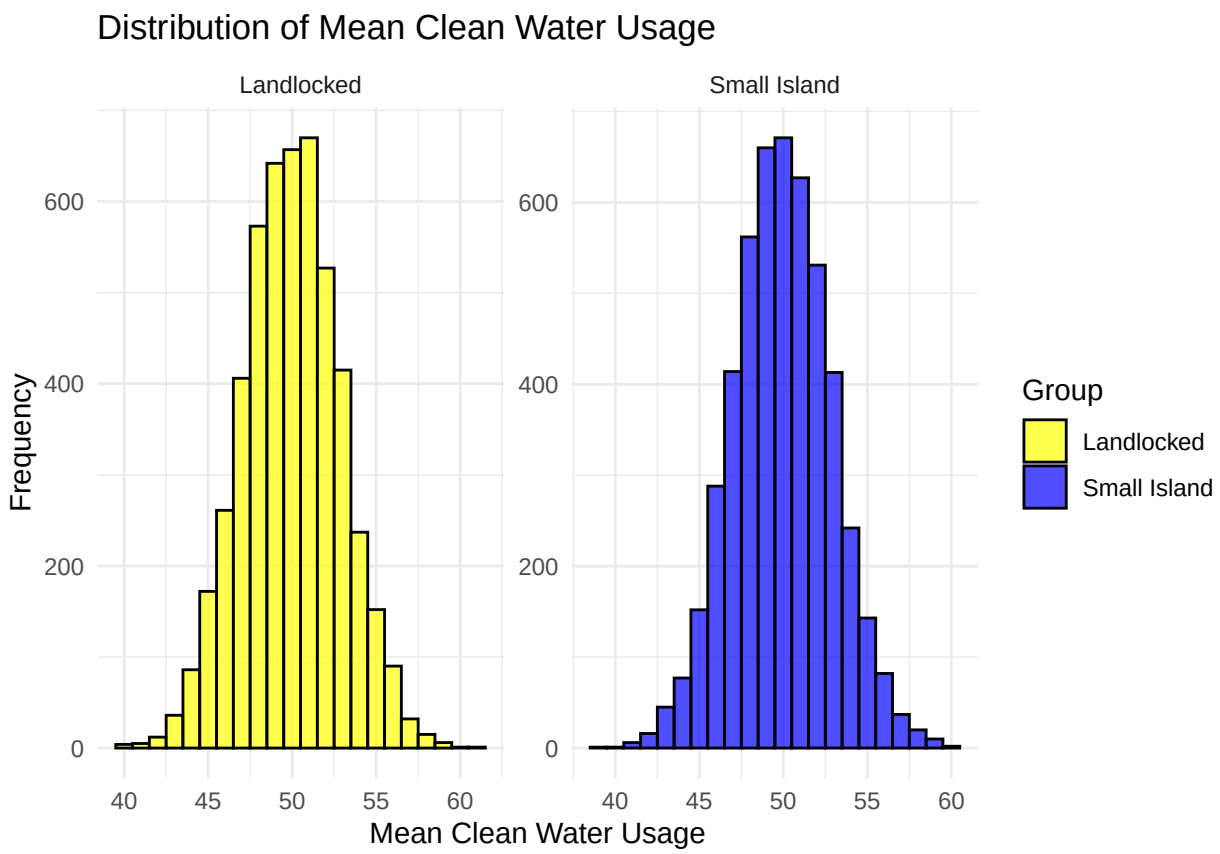
```
##                   76.24054                   90.57590
```

```
print(data3)
```

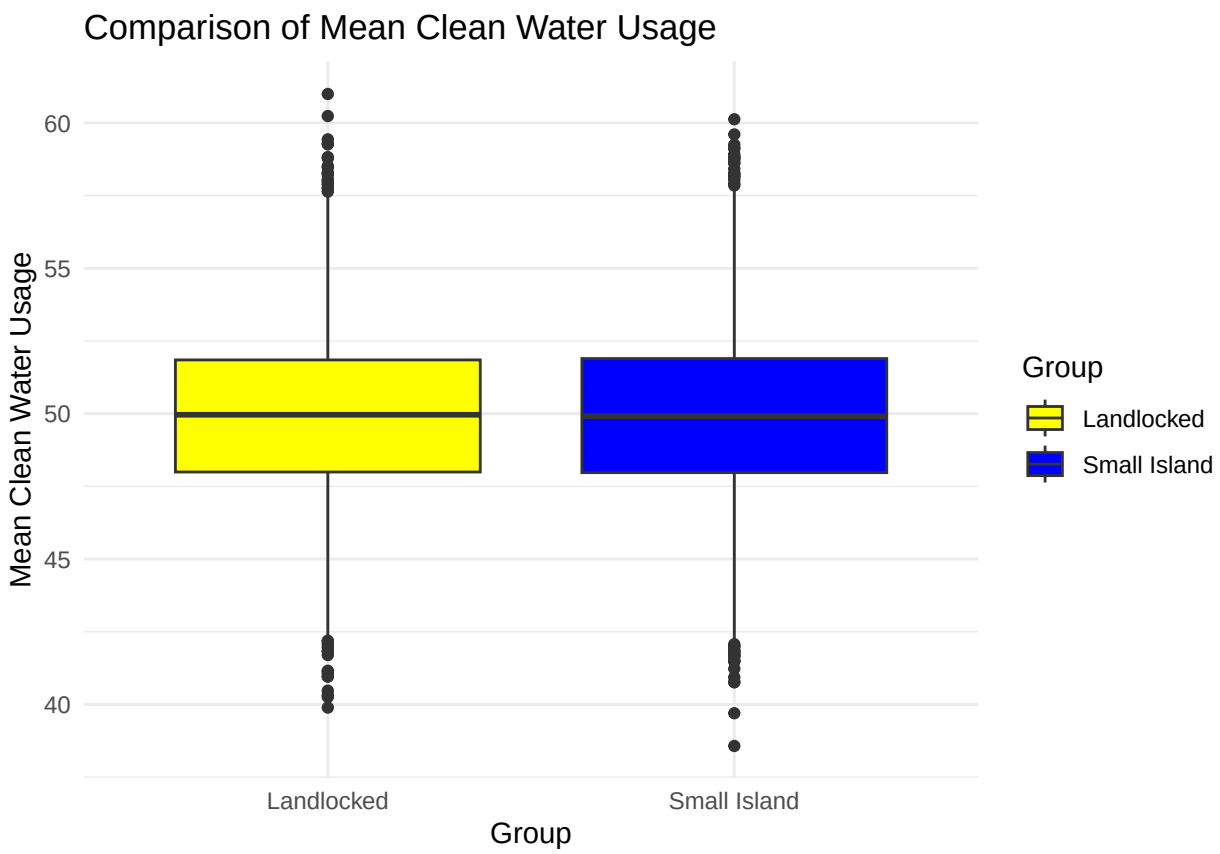
```
## # A tibble: 64 x 8
##    country_code_iso3 clean_water_usage_total clean_water_usage_urban
##    <chr>                               <dbl>                   <dbl>
##  1 AFG                                  75.1                   100
##  2 ARM                                 100.                    100.
##  3 AZE                                  96.0                   100
##  4 BRB                                  98.5                    NA
##  5 BLZ                                  98.4                    98.9
##  6 BTN                                  97.3                    98.1
##  7 BOL                                  93.4                    99.1
##  8 BWA                                  92.2                    97.6
##  9 VGB                                  99.9                    NA
## 10 BFA                                  47.2                    80.1
## # i 54 more rows
## # i 5 more variables: clean_water_usage_rural <dbl>,
## #   `Land Locked Developing Countries (LLDC) (M49)` <lgl>,
## #   `Small Island Developing States (SIDS) (M49)` <lgl>, goal_6_score <dbl>,
## #   Group <chr>
```

Visualizations

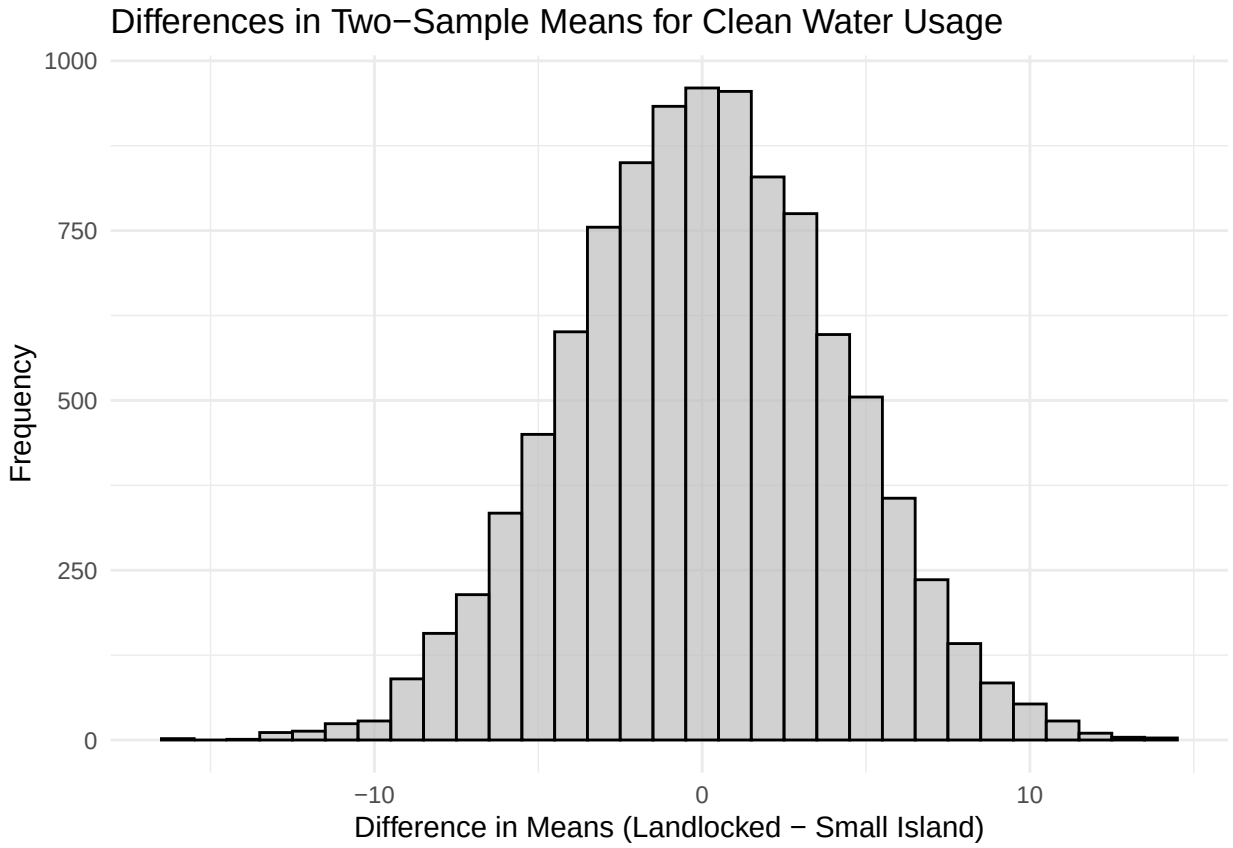
```
# Visualization of the mean clean water usage simulations using histograms
ggplot(data.frame(mean_clean_water_usage = clean_water_usage_simulations,
                  Group = rep(c("Landlocked", "Small Island"),
                              each = n_trials/2)),
       aes(x = mean_clean_water_usage, fill = Group)) +
  geom_histogram(binwidth = 1, position = "identity", alpha = 0.7, color = "black") +
  labs(title = "Distribution of Mean Clean Water Usage",
       x = "Mean Clean Water Usage",
       y = "Frequency") +
  facet_wrap(~ Group, scales = "free") +
  theme_minimal() +
  scale_fill_manual(values = c("Landlocked" = "yellow", "Small Island" = "blue"))
```




```
# Visualize using boxplots
ggplot(data.frame(mean_clean_water_usage = clean_water_usage_simulations,
                  Group = rep(c("Landlocked", "Small Island"),
                              each = n_trials/2)),
       aes(x = Group, y = mean_clean_water_usage, fill = Group)) +
  geom_boxplot() +
  labs(title = "Comparison of Mean Clean Water Usage",
       x = "Group",
       y = "Mean Clean Water Usage") +
  theme_minimal() +
  scale_fill_manual(values = c("Landlocked" = "yellow", "Small Island" = "blue"))
```




```
mean_difference_simulations <- numeric(n_trials)

# Simulate clean water usage for Landlocked countries
clean_water_landlocked <- matrix(runif(n_trials * n_sample, min = 0, max = 100), ncol = n_sample)

# Simulate clean water usage for Small Island countries
clean_water_small_island <- matrix(runif(n_trials * n_sample, min = 0, max = 100), ncol = n_sample)

for (i in 1:n_trials) {
  # Calculate mean clean water usage for Landlocked and Small Island countries
  mean_landlocked <- mean(clean_water_landlocked[i, ])
  mean_small_island <- mean(clean_water_small_island[i, ])
  mean_difference_simulations[i] <- mean_landlocked - mean_small_island
}

# Plotting histogram
ggplot(data.frame(mean_difference = mean_difference_simulations),
       aes(x = mean_difference)) +
  geom_histogram(binwidth = 1, color = "black", fill = "gray", alpha = 0.7) +
  labs(title = "Differences in Two-Sample Means for Clean Water Usage",
       x = "Difference in Means (Landlocked - Small Island)",
       y = "Frequency") +
  theme_minimal()
```



Research Question 3 ## Specific Research Question

Do countries show higher/lower gender inequality due to their geographic categorization as landlocked or small islands based on the mean Women Empowerment Index scores?

Data Wrangling and Cleaning

```
# load in country indicators with required variables
country_indicators <-
  read_csv("country_indicators.csv") %>%
  select(
    'iso3',
    'hdr_pr_f_2021',
    'sowc_maternal-and-newborn-health__demand-for-family-planning-satisfied-with-modern-methods-2016-20
    'sowc_adolescent-health__adolescent-birth-rate-2016-2021-r_aged-15-19_female',
    'sowc_women-s-economic-empowerment__educational-attainment-2008-2021-r_upper-secondary_female',
    'sowc_adolescents__transition-to-work-2013-2021-r_not-in-education-employment-or-training_female',
    'sowc_women-s-economic-empowerment__labour-force-participation-rate-2010-2020-r_female_total',
    'sowc_women-s-economic-empowerment__financial-inclusion-2014-2020-r_female_female',
    'sowc_adolescents__protection_intimate-partner-violence-2013-2020-r_female') %>%
  select(iso3, everything()) %>%
  rename(con_code = iso3)
```

```
## New names:
## Rows: 218 Columns: 1332
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
```

```
## (8): iso3, hdr_hdicode, hdr_region, wbi_income_group, wbi_lending_cat... dbl
## (1324): ...1, sowc_demographics__population-thousands-2021_total, sowc_d...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
# preview data
country_indicators
```

```
## # A tibble: 218 x 9
##    con_code hdr_pr_f_2021 sowc_maternal-and-newborn-hea~1 sowc_adolescent-heal~2
##    <chr>            <dbl>                          <dbl>                  <dbl>
##  1 AFG              27.2                            42.2                   62
##  2 ALB              35.7                             6.3                   14.2
##  3 DZA               7.48                           77.2                   12
##  4 AND              46.4                            NA                     2.7
##  5 AGO              29.5                            29.8                  163
##  6 AIA              NA                              NA                     40.1
##  7 ATG              31.4                            NA                     30.4
##  8 ARG              44.4                            NA                     40.9
##  9 ARM              33.6                            40.2                   18.9
## 10 AUS              37.9                            NA                     8.7
## # i 208 more rows
## # i abbreviated names:
## #   1: `sowc_maternal-and-newborn-health__demand-for-family-planning-satisfied-with-modern-methods-20
## #   2: `sowc_adolescent-health__adolescent-birth-rate-2016-2021-r_aged-15-19_female`
## # i 5 more variables:
## #   `sowc_women-s-economic-empowerment__educational-attainment-2008-2021-r_upper-secondary_female` <d
## #   `sowc_adolescents__transition-to-work-2013-2021-r_not-in-education-employment-or-training_female`
```

sdg file

```
# load SDG data and select necessary variable
sdg <-
  read_csv("sdr_fd5e4b5a.csv") %>%
  select('Country Code ISO3') %>% rename(con_code = 'Country Code ISO3')
```

```
## New names:
## Rows: 206 Columns: 59
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (36): Goal 1 Dash, Goal 1 Trend, Goal 2 Dash, Goal 2 Trend, Goal 3 Dash,... dbl
## (23): ...1, Goal 1 Score, Goal 2 Score, Goal 3 Score, Goal 4 Score, Goal...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
# preview data
sdg
```

```
## # A tibble: 206 x 1
##    con_code
##    <chr>
##  1 FIN
##  2 SWE
##  3 DNK
##  4 DEU
```

```
##  5 AUT
##  6 FRA
##  7 NOR
##  8 CZE
##  9 POL
## 10 EST
## # i 196 more rows
```

country_codes file

```
# load country_codes data and select variables
country_codes <-
  read_csv("country_codes.csv") %>%
  select('ISO-alpha3 Code (M49)','Small Island Developing States (SIDS) (M49)',
         'Land Locked Developing Countries (LLDC) (M49)') %>%
  rename(con_code = 'ISO-alpha3 Code (M49)')
```

```
## New names:
## Rows: 298 Columns: 125
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (99): Global Name_en (M49), Region Name_en (M49), Sub-region Name_en (M4... dbl
## (22): ...1, Global Code (M49), Region Code (M49), Intermediate Region Co... lgl
## (4): Sub-region Code (M49), Least Developed Countries (LDC) (M49), Land...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```

```
country_codes
```

```
## # A tibble: 298 x 3
##    con_code `Small Island Developing States (SIDS) (M49)` Land Locked Developi~1
##    <chr>    <lgl>                                         <lgl>
##  1 DZA      NA                                            NA
##  2 EGY      NA                                            NA
##  3 LBY      NA                                            NA
##  4 MAR      NA                                            NA
##  5 SDN      NA                                            NA
##  6 TUN      NA                                            NA
##  7 ESH      NA                                            NA
##  8 IOT      NA                                            NA
##  9 BDI      NA                                            TRUE
## 10 COM      TRUE                                          NA
## # i 288 more rows
## # i abbreviated name: 1: `Land Locked Developing Countries (LLDC) (M49)`
```

Data Integration

2- sample, 2-sided hypothesis testing

Null Hypothesis- There is no difference in the mean Women Empowerment Index scores between Landlocked and Small Islands H_0: $g_1 = g_2$ and $g_1 - g_2 = 0$

Alternate Hypothesis- There is a difference in the mean Women Empowerment Index scores between Landlocked and Small Islands H_1: $g_1 ! = g_2$ and $g_1 - g_2 ! = 0$

The set significance level (alpha level) would be set to $\alpha = 0.05$. Null hypothesis would be rejected if p <= α

```r
# Integrating data from all three data sets using con_code as a common key

wei_rough_data <- inner_join(x=country_codes, y=country_indicators, by="con_code")
wei_clean <- inner_join(x=wei_rough_data, y=sdg, by="con_code")

# Renaming columns and removing extra variables

wei_rename <- wei_clean %>%
  mutate(country_type = case_when(
      `Small Island Developing States (SIDS) (M49)` == TRUE ~ "Small Island",
      `Land Locked Developing Countries (LLDC) (M49)` == TRUE ~ "Land Locked",)) %>%
  rename(
    MMC ='sowc_maternal-and-newborn-health__demand-for-family-planning-satisfied-with-modern-methods-20
    ABR = 'sowc_adolescent-health__adolescent-birth-rate-2016-2021-r_aged-15-19_female',
    CSE = 'sowc_women-s-economic-empowerment__educational-attainment-2008-2021-r_upper-secondary_female
    NEET = 'sowc_adolescents__transition-to-work-2013-2021-r_not-in-education-employment-or-training_fe
    LFPR = 'sowc_women-s-economic-empowerment__labour-force-participation-rate-2010-2020-r_female_total
    FI = 'sowc_women-s-economic-empowerment__financial-inclusion-2014-2020-r_female_female',
    PR = hdr_pr_f_2021,
    IVP = 'sowc_adolescents__protection_intimate-partner-violence-2013-2020-r_female') %>%
  select(con_code, country_type, everything(), -"Land Locked Developing Countries (LLDC) (M49)",
         -"Small Island Developing States (SIDS) (M49)")

# replacing missing values of FI and IVP with their mean scores
wei_rename_fill <- wei_rename %>% mutate(
    FI = if_else(is.na(FI), mean(FI, na.rm = TRUE), FI),
    IVP = if_else(is.na(IVP), mean(IVP, na.rm = TRUE), IVP)) %>% drop_na()


wei_rename_fill
```

```
## # A tibble: 32 x 10
##    con_code country_type    PR   MMC   ABR   CSE  NEET  LFPR    FI   IVP
##    <chr>    <chr>        <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 BDI      Land Locked  38.9   39.6  58.2  3.84  10.9  80.5  6.73  37.8
##  2 ETH      Land Locked  39.5   63.6  73.5  5.66  17.7  73.3 29.1   24.3
##  3 MWI      Land Locked  22.9   73.9 102.   2.39  16.4  72.5 56.1   28.1
##  4 RWA      Land Locked  55.7   62.9  31.8  8.27  22.8  83.9 56.1   22.0
##  5 UGA      Land Locked  33.8   55.1 128.   6.35  18.1  66.9 56.1   31
##  6 ZMB      Land Locked  15.1   65.9 135    8.87  24.3  70.4 40.3   26.7
##  7 ZWE      Land Locked  34.6   84.8 108.   9.36  28.2  78.1 51.7   31.4
##  8 TCD      Land Locked  32.3   17.5 138.   1.84  44    64   56.1   14.5
##  9 LSO      Land Locked  22.9   82.8  90.8 14.2   30.3  60.4 46.5   22.0
## 10 BFA      Land Locked   6.30  52.6 124.   3.58  43.9  58.5 56.1    4.7
## # i 22 more rows
```

Creating Women Empowerment index-

```r
# Setting maximum and minimum values for each indicator taken from
# Technical Note: Twin Indices on Women's Empowerment and Gender Equality

min_MMC = 0
min_ABR = 0
```

```r
min_CSE = 0
min_NEET = 0
min_LFPR = 0
min_FI = 0
min_PR = 0
min_IVP = 0

max_MMC = 100
max_ABR = 200
max_CSE = 100
max_NEET = 85
max_LFPR = 100
max_FI = 100
max_PR = 75
max_IVP = 60

# Normalizing each variable to help with comparison
# All variables are positive indicators (higher values indicate better performance in
# that field) except ABR, NEET, IPV

wei_data <- wei_rename_fill %>%
  mutate(

    # Normalizing positive indicators
    norm_MMC  = (MMC - min_MMC)/(max_MMC - min_MMC),
    norm_FI   = (FI - min_FI)/(max_FI - min_FI),
    norm_PR   = (PR - min_PR)/(max_PR - min_PR),
    norm_CSE  = (CSE - min_CSE)/(max_CSE - min_CSE),
    norm_LFPR = (LFPR - min_LFPR)/(max_LFPR - min_LFPR),

    # Normalizing negative indicators
    norm_ABR  = (max_ABR - ABR)/(max_ABR - min_ABR),
    norm_NEET = (max_NEET - NEET)/(max_NEET - min_NEET),
    norm_IVP  = (max_IVP - IVP)/(max_IVP - min_IVP))

# Calculation of dimension indices
wei_data <- wei_data %>%
  mutate(
    I_health    = (norm_MMC + norm_ABR) / 2,
    I_education = (norm_CSE + norm_NEET) / 2,
    I_inclusion = (norm_LFPR + norm_FI) / 2,
    I_decision  = norm_PR,
    I_violence  = norm_IVP)


# Computing Women Empowerment Index (WEI is a positive index)
wei <- wei_data %>% select(-MMC, -ABR, -CSE, -NEET, -LFPR, -FI, -PR,-IVP) %>%
  mutate(WEI = (I_health * I_education * I_inclusion * I_decision * I_violence) ^ (1/5))

wei
```

```
## # A tibble: 32 x 16
##    con_code country_type norm_MMC norm_FI norm_PR norm_CSE norm_LFPR norm_ABR
##    <chr>    <chr>           <dbl>   <dbl>   <dbl>    <dbl>     <dbl>    <dbl>
```

```
##  1 BDI      Land Locked     0.396  0.0673  0.519    0.0384     0.805    0.709
##  2 ETH      Land Locked     0.636  0.291   0.527    0.0566     0.733    0.632
##  3 MWI      Land Locked     0.739  0.561   0.306    0.0239     0.725    0.491
##  4 RWA      Land Locked     0.629  0.561   0.742    0.0827     0.839    0.841
##  5 UGA      Land Locked     0.551  0.561   0.451    0.0635     0.669    0.360
##  6 ZMB      Land Locked     0.659  0.403   0.201    0.0887     0.704    0.325
##  7 ZWE      Land Locked     0.848  0.517   0.461    0.0936     0.781    0.460
##  8 TCD      Land Locked     0.175  0.561   0.430    0.0184     0.64     0.308
##  9 LSO      Land Locked     0.828  0.465   0.305    0.142      0.604    0.546
## 10 BFA      Land Locked     0.526  0.561   0.0840   0.0358     0.585    0.382
## # i 22 more rows
## # i 8 more variables: norm_NEET <dbl>, norm_IVP <dbl>, I_health <dbl>,
## #   I_education <dbl>, I_inclusion <dbl>, I_decision <dbl>, I_violence <dbl>,
## #   WEI <dbl>
```

Calculating the observed test statistic-

$\Delta\hat{g}$ will be the difference in the mean scores of WEI of landlocked and small island countries

```
ghat <- wei %>%
  group_by(country_type) %>%
  summarise(means = mean(WEI))
ghat
```

```
## # A tibble: 2 x 2
##   country_type means
##   <chr>        <dbl>
## 1 Land Locked  0.496
## 2 Small Island 0.399
```

```
delta_ghat <-
  wei %>%
  group_by(country_type) %>%
  summarise(means = mean(WEI)) %>%
  summarise(value = diff(means)) %>%
  as.numeric()

print(delta_ghat)
```

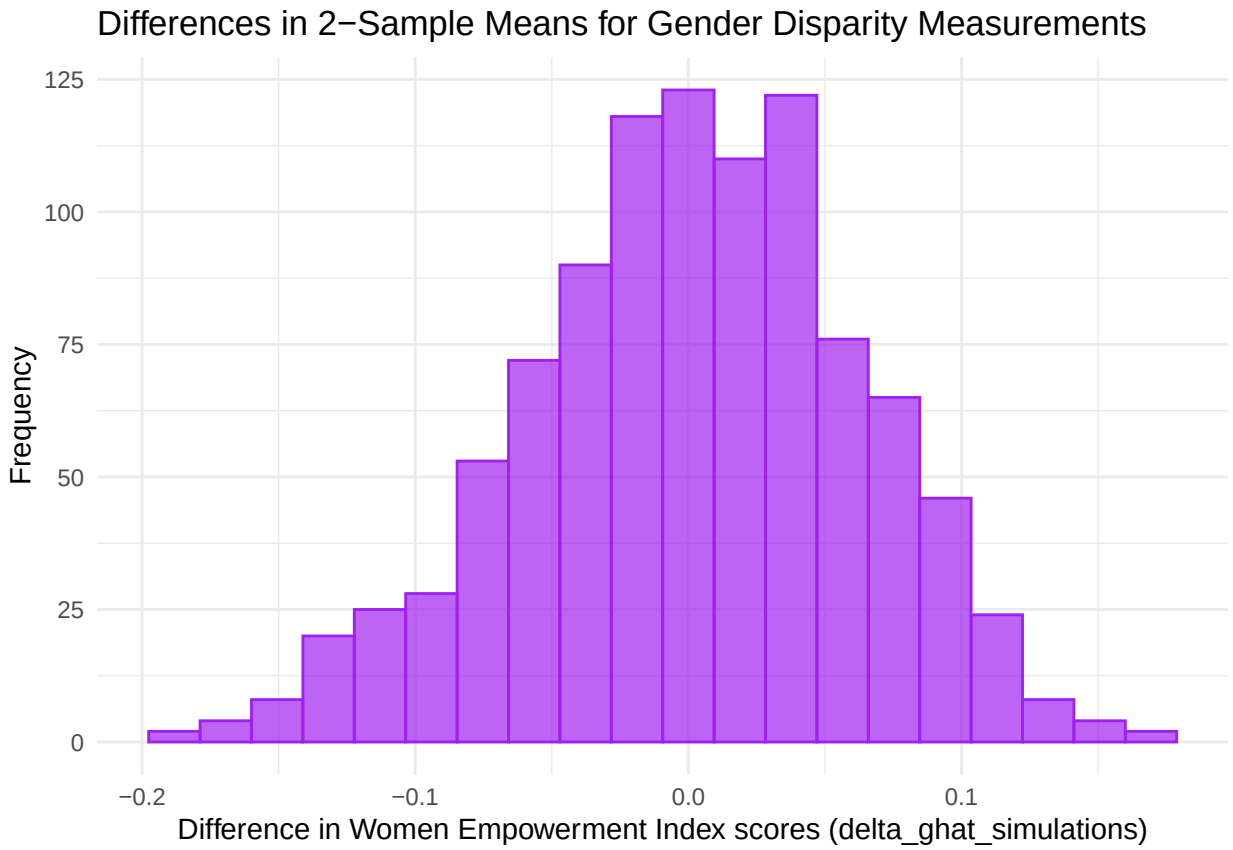
```
## [1] -0.09744326
```

Below is R code that simulates $N = 1000$ values of the test statistic $\Delta\hat{g}_{\text{sim}}$ **under the null hypothesis** using a permutation test. In this test, we assume that our groups are identical under our null hypothesis. Mixing the two groups together, randomly generating new groups with the same sizes, and then recomputing our test statistic each time therefore should allow us to simulate values from the sampling distribution provided our sample size is large enough.

```
seed_num <- 130
set.seed(seed_num)  # creating seed

# setup
n_trials <- 1000  # number of permutations

# simulating test statistic (difference between mean scores of WEI)
delta_ghat_simulations <- rep(NA, n_trials)

for(i in 1:n_trials){
  # perform a random permutation
  simdata <-
    wei %>%
    mutate(country_type = sample(country_type, replace=FALSE))

  # compute the simulated test statistic
  delta_ghat_sim <-
    simdata %>%
    group_by(country_type) %>%
    summarise(means = mean(WEI), .groups="drop") %>%
    summarise(value = diff(means)) %>%
    as.numeric()

  # store the simulated value
  delta_ghat_simulations[i] <- delta_ghat_sim
}
```

Visualizations-

```
# Visualizing sampling distribution of simulated test statistics using histograms

ggplot() +
  geom_histogram(aes(x=delta_ghat_simulations), color = "purple", fill = "purple",
                 position = "Identity", alpha = 0.7, bins= 20) +
  labs(x = "Difference in Women Empowerment Index scores (delta_ghat_simulations)",
       y = "Frequency",
       title = "Differences in 2-Sample Means for Gender Disparity Measurements") +
  theme_minimal() +
  theme(legend.position = "top")
```




```
# Creating box plots to compare the median WEI scores

ggplot(data = wei, aes(x = country_type, y = WEI, fill = country_type)) +
  geom_boxplot() +
  labs(x = "Country Type", y = "Women Empowerment Index Scores",
       title = "Comparison of Women Empowerment Index Scores by Country Type") +
  scale_fill_manual(values = c("orange", "purple")) +
  theme_minimal() +
  theme(legend.position = "top")
```

Computing the p-value-

(the probability of observing a test statistic at least as extreme as the observed value if the null hypothesis is true)

```
# null hypothesis value
delta_median_null <- 0

p_value <-sum(abs(delta_ghat_simulations - delta_median_null) >=
                abs(delta_ghat - delta_median_null)) / n_trials
print(p_value)
```

```
## [1] 0.115
```

Citations (MLA 9th edition)

1. Jain-Chandra, Sonali. "Chapter 2. Gender Inequality around the World." Www.elibrary.imf.org, International Monetary Fund, www.elibrary.imf.org/display/book/9781513516103/ch002.xml#:~:text=The%20gender%20gap%20varies%20strongly.
2. TOWARDS IMPROVED MEASURES of GENDER INEQUALITY: An Evaluation of the UNDP Gender Inequality Index and a Proposal. www.unwomen.org/sites/default/files/2022-11/Discussion-paper-Towards-improved-measures-of-gender-inequality-en.pdf.
3. Technical Note: Twin Indices on Women's Empowerment and Gender Equality. hdr.undp.org/sites/default/files/publications/additional-files/2023-07/paths_equal_2023_tn.pdf.

4. Duflo, Esther. "Women Empowerment and Economic Development." *Journal of Economic Literature*, vol. 50, no. 4, 2012, pp. 1051-79.