Tidy Tuesday

Week 39

AUTHOR
Cristian T

PUBLISHED
October 6, 2024

This week we are exploring the International Mathematical Olympiad (IMO)!

The International Mathematical Olympiad (IMO) is the World Championship Mathematics Competition for High School students and is held annually in a different country. The first IMO was held in 1959 in Romania, with 7 countries participating. It has gradually expanded to over 100 countries from 5 continents. The competition consists of 6 problems and is held over two consecutive days with 3 problems each.

`country_results_df.csv`

variable	class	description
year	integer	Year of IMO
country	character	Participating country
team_size_all	integer	Participating contestants
team_size_male	integer	Male contestants
team_size_female	integer	Female contestants
p1	integer	Score on problem 1
p2	integer	Score on problem 2
p3	integer	Score on problem 3
p4	integer	Score on problem 4
p5	integer	Score on problem 5
p6	integer	Score on problem 6
p7	integer	Score on problem 7
awards_gold	integer	Number of gold medals
awards_silver	integer	Number of silver medals
awards_bronze	integer	Number of bronze medals
awards_honorable_mentions	integer	Number of honorable mentions
leader	character	Leader of country team
deputy_leader	character	Deputy leader of country team

`timeline_df.csv`

variable	class	description
year	integer	Year of IMO
contestant	character	Participant’s name
country	character	Participant’s country
p1	integer	Score on problem 1
p2	integer	Score on problem 2
p3	integer	Score on problem 3
p4	integer	Score on problem 4
p5	integer	Score on problem 5
p6	integer	Score on problem 6
total	integer	Total score on all problems
individual_rank	integer	Individual rank
award	character	Award won

`timeline_df.csv`

variable	class	description
edition	integer	Edition of International Mathematical Olympiad (IMO)
year	integer	Year of IMO
country	character	Host country
city	character	Host city
countries	integer	Number of participating countries
all_contestant	integer	Number of participating contestants
male_contestant	integer	Number of participating male contestants
female_contestant	integer	Number of participating female contestants
start_date	Date	Start date of IMO
end_date	Date	End date of IMO

Load the data

```
        # Load the tidytuesday package
        suppressMessages(library(tidytuesdayR)) # For accessing TidyTuesday datasets
        suppressMessages(library(skimr)) # For summary and descriptive statistics
        suppressMessages(library(tidyverse)) # For data manipulation and visualization
        suppressMessages(library(dplyr)) # For data manipulation and transformation
        suppressMessages(library(ggplot2)) # For data visualization
        suppressMessages(library(RColorBrewer)) # For color palettes in visualizations
        suppressMessages(library(ggimage)) # For adding images to plots
        suppressMessages(library(tidytext))
        suppressMessages(library(sentimentr))
        suppressMessages(library(ggpubr))


        # Load the current week's dataset
        tuesdata <- tidytuesdayR::tt_load('2024-09-24')
```

```
    Downloading file 1 of 3: `country_results_df.csv`
    Downloading file 2 of 3: `individual_results_df.csv`
    Downloading file 3 of 3: `timeline_df.csv`
```

```
        # Extract datasets from the TidyTuesday dataset
        country_results <- tuesdata$country_results_df
        individual_results <- tuesdata$individual_results_df
        timeline <- tuesdata$timeline_df

        # Rename datasets
        #ca <- college_admissions


        # Explore the structure of the dataset
        str(country_results) # Display the structure of 'hamlet'
```

```
spc_tbl_ [3,780 × 18] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ year                      : num [1:3780] 2024 2024 2024 2024 2024 ...
 $ country                   : chr [1:3780] "United States of America" "People's Republic of
China" "Republic of Korea" "India" ...
 $ team_size_all             : num [1:3780] 6 6 6 6 6 6 6 6 6 6 ...
 $ team_size_male            : num [1:3780] 5 6 6 6 6 6 6 6 6 5 ...
 $ team_size_female          : num [1:3780] 1 0 0 0 0 0 0 0 0 1 ...
 $ p1                        : num [1:3780] 42 42 42 42 42 42 42 42 42 38 ...
 $ p2                        : num [1:3780] 41 42 37 34 30 37 33 37 25 37 ...
 $ p3                        : num [1:3780] 19 31 18 11 10 7 8 16 5 5 ...
 $ p4                        : num [1:3780] 40 40 42 42 42 42 42 36 42 42 ...
 $ p5                        : num [1:3780] 35 22 7 28 36 29 31 23 35 12 ...
 $ p6                        : num [1:3780] 15 13 22 10 5 5 6 1 2 17 ...
 $ p7                        : logi [1:3780] NA NA NA NA NA NA ...
 $ awards_gold               : num [1:3780] 5 5 2 4 4 1 2 2 1 2 ...
 $ awards_silver             : num [1:3780] 1 1 4 1 0 5 3 3 4 2 ...
 $ awards_bronze             : num [1:3780] 0 0 0 0 2 0 1 1 1 2 ...
 $ awards_honorable_mentions: num [1:3780] 0 0 0 1 0 0 0 0 0 0 ...
 $ leader                    : chr [1:3780] "John Berman" "Liang Xiao" "Suyoung Choi"
"Krishnan Sivasubramanian" ...
```

```
 $ deputy_leader            : chr [1:3780] "Carl Schildkraut" "Yijun Yao" "Hwajong Yoo"
"Rijul Saini" ...
 - attr(*, "spec")=
  .. cols(
  ..   year = col_double(),
  ..   country = col_character(),
  ..   team_size_all = col_double(),
  ..   team_size_male = col_double(),
  ..   team_size_female = col_double(),
  ..   p1 = col_double(),
  ..   p2 = col_double(),
  ..   p3 = col_double(),
  ..   p4 = col_double(),
  ..   p5 = col_double(),
  ..   p6 = col_double(),
  ..   p7 = col_logical(),
  ..   awards_gold = col_double(),
  ..   awards_silver = col_double(),
  ..   awards_bronze = col_double(),
  ..   awards_honorable_mentions = col_double(),
  ..   leader = col_character(),
  ..   deputy_leader = col_character()
  .. )
 - attr(*, "problems")=<externalptr>
```

```
        str(individual_results) # Display the structure of 'macbeth'
```

```
spc_tbl_ [21,707 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ year           : num [1:21707] 2024 2024 2024 2024 2024 ...
 $ contestant     : chr [1:21707] "Haojia Shi" "Ivan Chasovskikh" "Alexander Wang" "Satoshi
Kano" ...
 $ country        : chr [1:21707] "People's Republic of China" "C21" "United States of
America" "Japan" ...
 $ p1             : num [1:21707] 7 7 7 7 7 7 7 7 7 7 ...
 $ p2             : num [1:21707] 7 7 7 7 7 7 7 2 7 7 ...
 $ p3             : num [1:21707] 7 6 3 2 7 4 7 7 2 5 ...
 $ p4             : num [1:21707] 7 6 7 7 7 7 7 7 7 7 ...
 $ p5             : num [1:21707] 7 7 7 7 7 7 7 7 7 7 ...
 $ p6             : num [1:21707] 7 7 7 7 0 3 0 5 5 2 ...
 $ p7             : logi [1:21707] NA NA NA NA NA NA ...
 $ total          : num [1:21707] 42 40 38 37 35 35 35 35 35 35 ...
 $ individual_rank: num [1:21707] 1 2 3 4 5 5 5 5 5 5 ...
 $ award          : chr [1:21707] "Gold medal" "Gold medal" "Gold medal" "Gold medal" ...
 - attr(*, "spec")=
  .. cols(
  ..   year = col_double(),
  ..   contestant = col_character(),
  ..   country = col_character(),
  ..   p1 = col_double(),
  ..   p2 = col_double(),
  ..   p3 = col_double(),
  ..   p4 = col_double(),
  ..   p5 = col_double(),
  ..   p6 = col_double(),
  ..   p7 = col_logical(),
  ..   total = col_double(),
  ..   individual_rank = col_double(),
  ..   award = col_character()
  .. )
 - attr(*, "problems")=<externalptr>
```

```
str(timeline) # Display the structure of 'romeo_juliet'
```

```
spc_tbl_ [65 × 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ edition           : num [1:65] 65 64 63 62 61 60 59 58 57 56 ...
 $ year              : num [1:65] 2024 2023 2022 2021 2020 ...
 $ country           : chr [1:65] "United Kingdom" "Japan" "Norway" "Russian Federation" ...
 $ city              : chr [1:65] "Bath" "Chiba" "Oslo" "A distributed IMO administered from
St Petersburg" ...
 $ countries         : num [1:65] 108 112 104 107 105 112 107 111 109 104 ...
 $ all_contestant    : num [1:65] 609 618 589 619 616 621 594 615 602 577 ...
 $ male_contestant   : num [1:65] 528 550 521 555 560 556 535 553 531 525 ...
 $ female_contestant: num [1:65] 81 67 68 64 56 65 59 62 71 52 ...
 $ start_date        : Date[1:65], format: "2024-07-11" "2023-07-02" ...
 $ end_date          : Date[1:65], format: "2024-07-22" "2023-07-13" ...
 - attr(*, "spec")=
  .. cols(
  ..   edition = col_double(),
  ..   year = col_double(),
  ..   country = col_character(),
  ..   city = col_character(),
  ..   countries = col_double(),
  ..   all_contestant = col_double(),
  ..   male_contestant = col_double(),
  ..   female_contestant = col_double(),
  ..   start_date = col_date(format = ""),
  ..   end_date = col_date(format = "")
  .. )
 - attr(*, "problems")=<externalptr>
```

```
skim(country_results) # Provide detailed summary statistics for 'hamlet' (missing val
```

Name	country_results
Number of rows	3780
Number of columns	18
_______________________	
Column type frequency:	
character	3
logical	1
numeric	14
________________________	
Group variables	None

Data summary

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
country	0	1.00	4	37	0	139	0
leader	870	0.77	6	38	0	997	0
deputy_leader	968	0.74	6	55	0	1383	0

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
p7	3780	0	NaN	:

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
year	0	1.00	2003.91	14.49	1959	1995.00	2006	2016	2024	▁▂▃▇▇
team_size_all	0	1.00	5.74	1.22	1	6.00	6	6	8	▁▁▁▇▁
team_size_male	283	0.93	5.20	1.46	0	5.00	6	6	8	▁▁▁▇▁
team_size_female	2180	0.42	1.07	0.77	0	1.00	1	1	6	▇▂▁▁▁
p1	110	0.97	24.74	14.05	0	12.00	26	38	56	▅▅▅▇▁
p2	110	0.97	15.44	13.48	0	3.25	12	26	56	▇▃▂▂▁
p3	110	0.97	6.96	10.38	0	0.00	2	10	64	▇▁▁▁▁
p4	110	0.97	23.01	14.06	0	10.00	23	36	56	▇▅▅▇▁
p5	110	0.97	14.09	13.11	0	2.00	10	23	56	▇▃▂▂▁
p6	110	0.97	5.70	9.84	0	0.00	1	7	63	▇▁▁▁▁
awards_gold	2	1.00	0.47	1.05	0	0.00	0	0	6	▇▁▁▁▁
awards_silver	2	1.00	0.96	1.29	0	0.00	0	2	6	▇▂▁▁▁
awards_bronze	2	1.00	1.42	1.37	0	0.00	1	2	6	▇▂▂▁▁
awards_honorable_mentions	515	0.86	1.18	1.34	0	0.00	1	2	6	▇▂▁▁▁

```r
        skim(individual_results) # Provide detailed summary statistics for 'hamlet' (missing
```

Name	individual_results
Number of rows	21707
Number of columns	13
_______________________	
Column type frequency:	
character	3
logical	1
numeric	9

________________________	
Group variables	None

Data summary

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
contestant	0	1.00	1	72	0	14721	0
country	0	1.00	3	37	0	157	0
award	7044	0.68	10	31	0	12	0

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
p7	21692	0	0.2	FAL: 12, TRU: 3

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
year	0	1.00	2003.01	15.31	1959	1994	2006	2016	2024	▁▂▅▇▇
p1	1093	0.95	4.36	2.90	0	1	6	7	8	▅▂▁▂▇
p2	1093	0.95	2.71	2.90	0	0	1	7	8	▇▂▁▁▃
p3	1093	0.95	1.21	2.21	0	0	0	1	9	▇▁▁▁▁
p4	1093	0.95	4.06	2.95	0	1	5	7	7	▆▁▂▁▇
p5	1093	0.95	2.46	2.85	0	0	1	6	8	▇▂▁▁▃
p6	1093	0.95	0.97	2.05	0	0	0	1	9	▇▁▁▁▁
total	0	1.00	15.98	10.47	0	8	15	23	46	▇▇▆▃▁
individual_rank	29	1.00	219.88	157.65	1	82	189	343	604	▇▆▅▃▂

```
        skim(timeline) # Provide detailed summary statistics for 'hamlet' (missing values, su
```

Name	timeline
Number of rows	65
Number of columns	10
_______________________	
Column type frequency:	
character	2
Date	2
numeric	6
________________________	
Group variables	None

Data summary

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
country	0	1	4	35	0	39	0
city	0	1	4	49	0	55	0

Variable type: Date

skim_variable	n_missing	complete_rate	min	max	median	n_unique
start_date	0	1	1959-07-21	2024-07-11	1992-07-10	65
end_date	0	1	1959-07-31	2024-07-22	1992-07-21	65

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
edition	0	1.00	33.00	18.91	1	17	33	49.0	65	▇▇▇▇▇
year	0	1.00	1991.68	19.29	1959	1975	1992	2008.0	2024	▇▇▇▇▇
countries	0	1.00	58.02	38.14	5	17	56	95.0	112	▇▁▁▅▇
all_contestant	0	1.00	333.69	202.85	39	132	322	522.0	621	▇▂▁▅▇
male_contestant	0	1.00	279.88	182.14	32	112	277	471.0	560	▇▂▅▁▇
female_contestant	1	0.98	26.70	23.98	1	4	21	49.5	81	▇▁▁▂▁

```r
# Export data
# write.csv(country_results, "country_results.csv", row.names = FALSE)
# write.csv(individual_results, "individual_results.csv", row.names = FALSE)
# write.csv(timeline, "timeline.csv", row.names = FALSE)




#write.csv(combined_plays, "combined_plays.csv", row.names = FALSE)

#tidytuesdayR::use_tidytemplate()

# country_results <- country_results %>%
#   left_join(standardized_countries, by = "country")
```

Standarized Countries

```r
# Load necessary libraries
library(dplyr)
library(countrycode)

# Create a new data frame with standardized country names
standardized_countries <- data.frame(
  country = unique(country_results$country),
  standardized_country = countrycode(unique(country_results$country), "country.name",
  stringsAsFactors = FALSE
)
```

```
# View the new standardized countries data frame
standardized_countries
```

country	standardized_country
<chr>	<chr>
United States of America	United States
People's Republic of China	China
Republic of Korea	South Korea
India	India
Belarus	Belarus
Singapore	Singapore
United Kingdom	United Kingdom
Hungary	Hungary
Poland	Poland
Türkiye	Turkey

1-10 of 139 rows Previous **1** 2 3 4 5 6 ... 14 Next

```
# Merge the new data frame with the original country_results
country_results <- country_results %>%
  left_join(standardized_countries, by = "country")

# View the updated country_results
country_results
```

year	country	team_size_all	team_size_male
<dbl>	<chr>	<dbl>	<dbl>
2024	United States of America	6	5
2024	People's Republic of China	6	6
2024	Republic of Korea	6	6
2024	India	6	6
2024	Belarus	6	6
2024	Singapore	6	6
2024	United Kingdom	6	6
2024	Hungary	6	6
2024	Poland	6	6
2024	Türkiye	6	5

1-10 of 3,780 rows | 1-4 of 19 columns Previous **1** 2 3 4 5 6 ... 378 Next

country_results2

```
# Create a copy of country_results
country_results2 <- country_results


# Create new columns for total males and females
country_results2 <- country_results2 %>%
  mutate(total_males = team_size_male,total_females = team_size_all - team_size_male)
```

```r
    mutate(female_proportion = total_females / team_size_all) %>%
      mutate(has_female = ifelse(total_females > 0, 1, 0))

  country_results2 %>%
    select(year, standardized_country, team_size_all, team_size_male, team_size_female,
```

year	standardized_country	team_size_all	team_size_male
<dbl>	<chr>	<dbl>	<dbl>
2024	United States	6	5
2024	China	6	6
2024	South Korea	6	6
2024	India	6	6
2024	Belarus	6	6
2024	Singapore	6	6
2024	United Kingdom	6	6
2024	Hungary	6	6
2024	Poland	6	6
2024	Turkey	6	5

1-10 of 3,780 rows | 1-4 of 8 columns Previous **1** 2 3 4 5 6 ... 378 Next

Merged GII Data 2022

The Gender Inequality Index (GII) is a composite measure that reflects inequalities in gender-based outcomes across three key dimensions: reproductive health, empowerment, and labor market participation.

```r
  # Create a copy of country_results
  country_results_gii <- country_results


  # Merge the datasets on the 'country' column
  country_results_gii <- country_results_gii %>%
    left_join(gii_data, c("standardized_country" = "country"))


  country_results_filtered <- country_results_gii %>%
    select(year, country, standardized_country, hdi.rank, value, rank)

  na <- country_results_filtered %>%
    filter(is.na(rank)) %>%
    group_by(country,standardized_country ) %>%  # Group by the country column
    summarise(
      count = n(),  # Count of rows with NA rank for each country
      .groups = 'drop'  # Optionally, drop the grouping after summarizing
    )
  na
```

country	standardized_country	count
<chr>	<chr>	<int>
Commonwealth of Independent States	NA	1
Czechoslovakia	Czechoslovakia	33

country <chr>	standardized_country <chr>	count <int>
German Democratic Republic	German Democratic Republic	29
Hong Kong	Hong Kong SAR China	37
Ivory Coast	Côte d'Ivoire	7
Kosovo	Kosovo	13
Liechtenstein	Liechtenstein	15
Macau	Macao SAR China	33
Palestine	Palestinian Territories	1
Puerto Rico	Puerto Rico	25

1-10 of 18 rows Previous 1 2 Next

```
# Create new columns for total males and females
country_results_gii <- country_results_gii %>%
  filter(year == 2022) %>%
  mutate(total_males = team_size_male,total_females = team_size_all - team_size_male)
  mutate(female_proportion = total_females / team_size_all) %>%
    mutate(has_female = ifelse(total_females > 0, 1, 0))

country_results_gii %>%
  select(year, standardized_country, team_size_all, team_size_male, team_size_female,
```

year <dbl>	standardized_country <chr>	team_size_all <dbl>	team_size_male <dbl>
2022	China	6	6
2022	South Korea	6	6
2022	United States	6	6
2022	Vietnam	6	6
2022	Romania	6	6
2022	Thailand	6	6
2022	Germany	6	6
2022	Iran	6	6
2022	Japan	6	6
2022	Israel	6	5

1-10 of 104 rows | 1-4 of 8 columns Previous 1 2 3 4 5 6 ... 11 Next

Merge GDI Data 1990 - 2022

The Gender Development Index (GDI)

```
# Create a copy of country_results
country_results_gdi <- country_results

# Merge the datasets on the 'country' column
country_results_gdi <- country_results %>%
  left_join(gdi_data, by = c("standardized_country" = "country", "year" = "year"))

# Remove rows with NA
```

```
country_results_gdi <- country_results_gdi %>%
  drop_na(gender.development.index)

country_results_filtered <- country_results_gdi %>%
  select(year, country, standardized_country, year, gender.development.index)

# Create new columns for total males and females
country_results_gdi <- country_results_gdi %>%
  mutate(total_males = team_size_male,total_females = team_size_all - team_size_male)
  mutate(female_proportion = total_females / team_size_all) %>%
    mutate(has_female = ifelse(total_females > 0, 1, 0))

country_results_gdi %>%
  select(year, standardized_country, team_size_all, team_size_male, team_size_female,
```

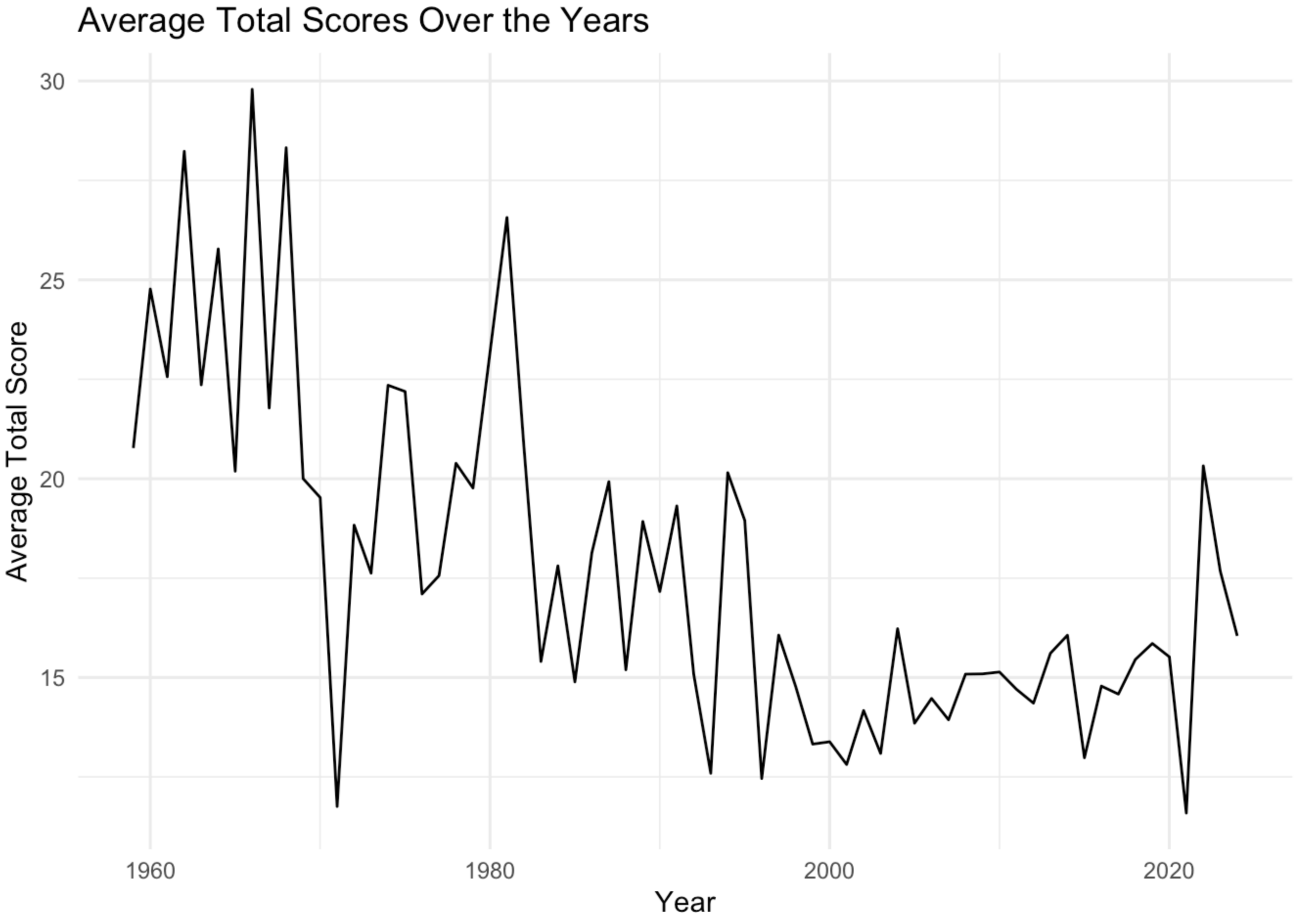
year	standardized_country	team_size_all	team_size_male
<dbl>	<chr>	<dbl>	<dbl>
2022	China	6	6
2022	South Korea	6	6
2022	United States	6	6
2022	Vietnam	6	6
2022	Romania	6	6
2022	Thailand	6	6
2022	Germany	6	6
2022	Iran	6	6
2022	Japan	6	6
2022	Israel	6	5

1-10 of 2,557 rows | 1-4 of 8 columns Previous **1** 2 3 4 5 6 ... 256 Next

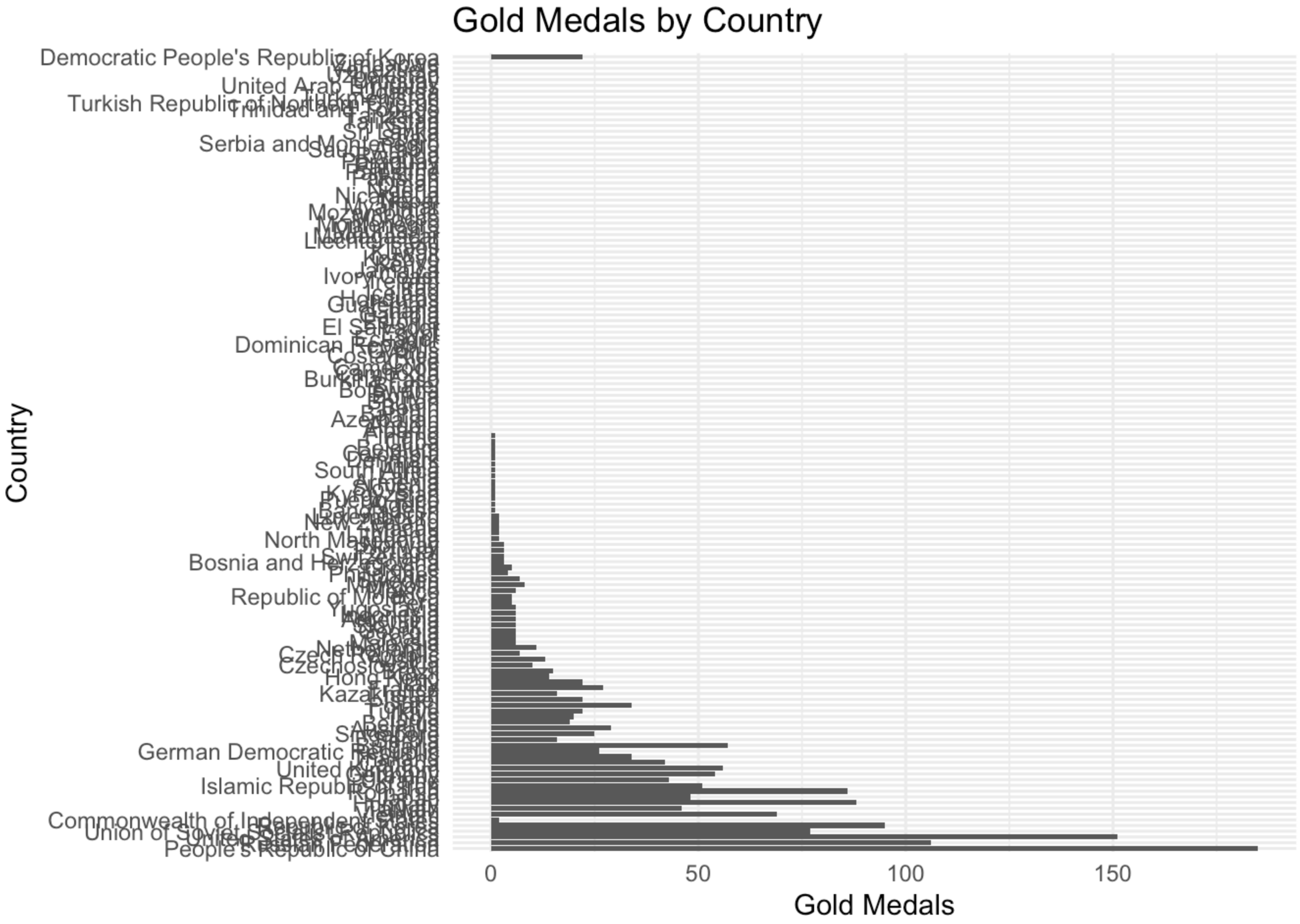
```
#### Clean the data

# Missing values are associated with line numbers and stage direction
```

EDA

```
# Trend of total scores over the years
ggplot(individual_results, aes(x = year, y = total)) +
    geom_line(stat = "summary", fun = mean) +
    labs(title = "Average Total Scores Over the Years",
         x = "Year",
         y = "Average Total Score") +
    theme_minimal()
```



```r
medal_distribution <- country_results %>%
    group_by(country) %>%
    summarize(total_gold = sum(awards_gold, na.rm = TRUE)) %>%
    arrange(desc(total_gold))

# Medal distribution by country
ggplot(country_results, aes(x = reorder(country, -awards_gold), y = awards_gold)) +
    geom_bar(stat = "identity") +
    labs(title = "Gold Medals by Country",
         x = "Country",
         y = "Gold Medals") +
    coord_flip() +
    theme_minimal()
```



```
#unique(country_results$country)
```

Plot 1 Female Participation Over Time

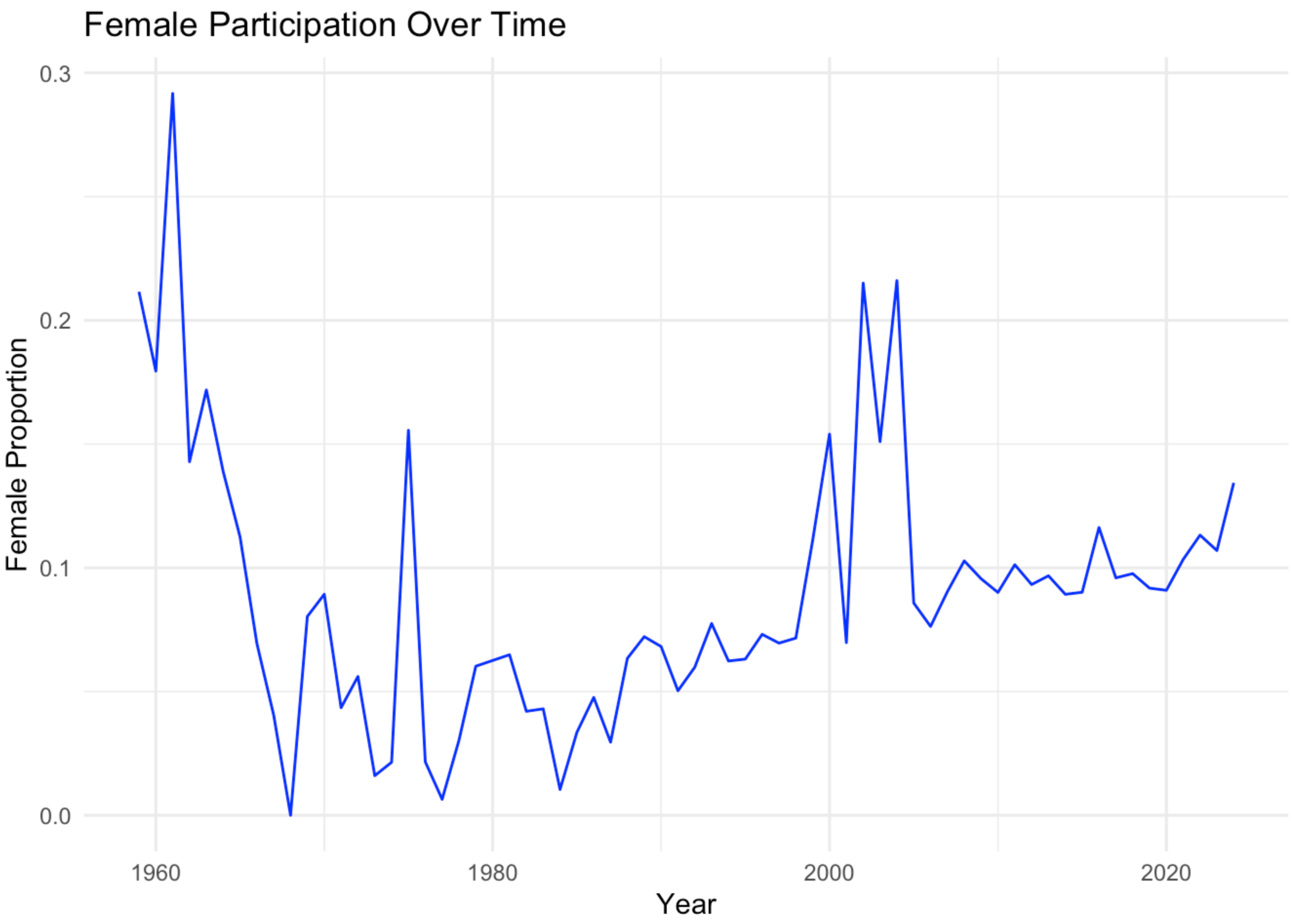
```
############ female participation over time by year

# Aggregate data by year
yearly_female_participation <- country_results2 %>%
  group_by(year) %>%
  summarize(
    total_females = sum(total_females, na.rm = TRUE),
    total_team_size = sum(team_size_all, na.rm = TRUE)
  ) %>%
  mutate(female_proportion = total_females / total_team_size)

# View the summary
yearly_female_participation %>% arrange(desc(female_proportion))
```

year	total_females	total_team_size	female_proportion
<dbl>	<dbl>	<dbl>	<dbl>
1961	14	48	0.291666667
2004	105	486	0.216049383
2002	103	479	0.215031315
1959	11	52	0.211538462
1960	7	39	0.179487179

year	total_females	total_team_size	female_proportion
<dbl>	<dbl>	<dbl>	<dbl>
1963	11	64	0.171875000
1975	21	135	0.155555556
2000	71	461	0.154013015
2003	69	457	0.150984683
1962	8	56	0.142857143

1-10 of 65 rows   Previous **1** 2 3 4 5 6 7 Next

```
ggplot(yearly_female_participation, aes(x = year, y = female_proportion)) +
  geom_line(color = "blue") +
  labs(title = "Female Participation Over Time",
       x = "Year",
       y = "Female Proportion") +
  theme_minimal()
```

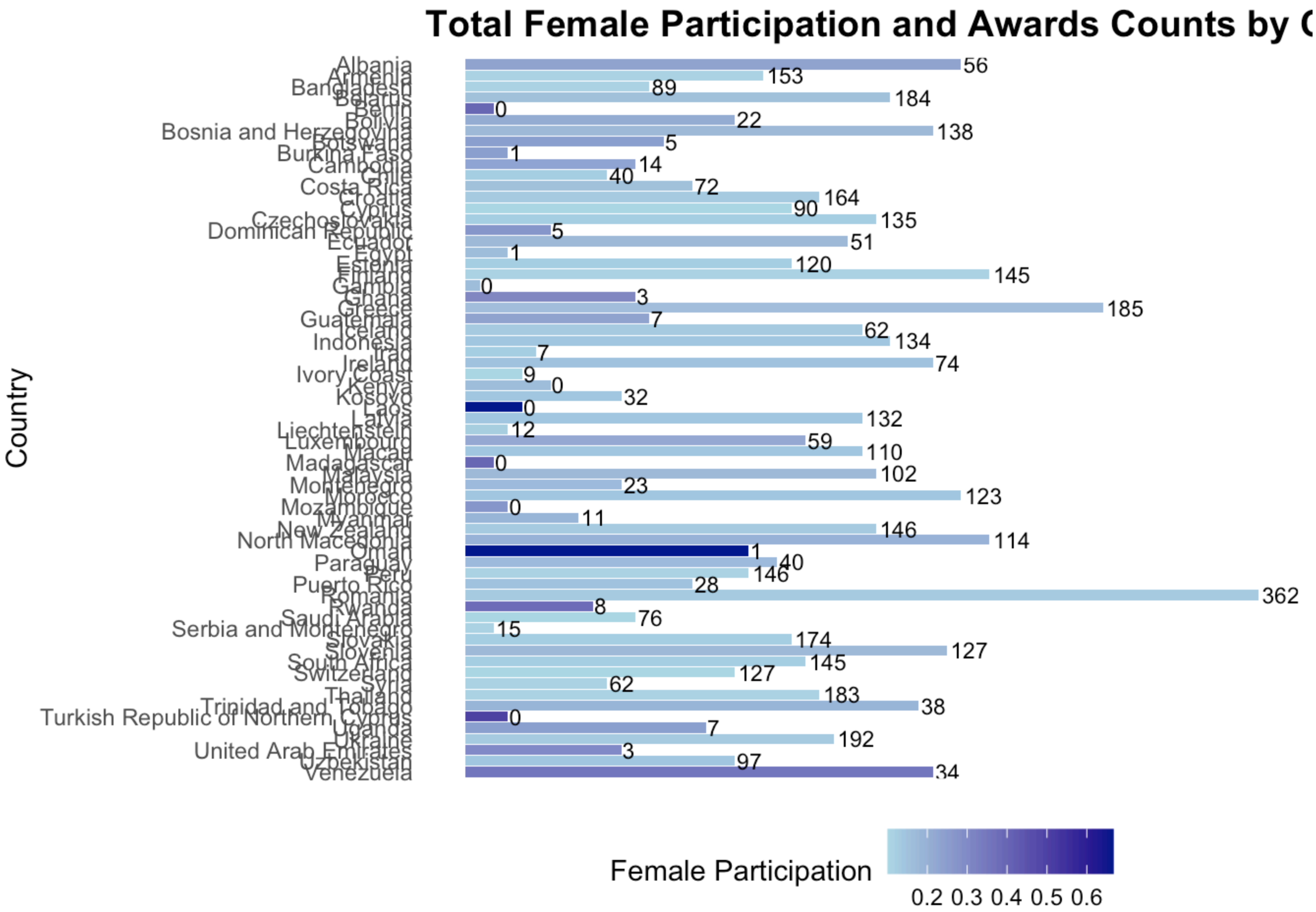


```
# country_results %>% filter(year == 1961)
```

Plot 2 Total Female Participation and Awards Counts by Country (1959 - 2024) with Over 10% Female Participation

```
############ female participation by country
country_female_participation2 <- country_results2 %>%
```

```r
  group_by(country) %>%
  summarize(
    total_females = sum(total_females, na.rm = TRUE),
    total_team_size = sum(team_size_all, na.rm = TRUE),
    total_awards_gold = sum(awards_gold, na.rm = TRUE),
    total_awards_silver = sum(awards_silver, na.rm = TRUE),
    total_awards_bronze = sum(awards_bronze, na.rm = TRUE),
    total_awards_honorable_mentions = sum(awards_honorable_mentions, na.rm = TRUE)
  ) %>%
  mutate(female_proportion = total_females / total_team_size)

# United States shows female participation 0.01923077

# Filter for > 10% female participation
country_female_participation2_filtered <- country_female_participation2 %>%
  filter(female_proportion > 0.1)

# Create a new variable to represent the total awards
country_female_participation2_filtered <- country_female_participation2_filtered %>%
  mutate(total_awards = total_awards_gold + total_awards_silver + total_awards_bronze


# Create the plot
ggplot(country_female_participation2_filtered, aes(x = fct_rev(country), y = total_fe
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Total Female Participation and Awards Counts by Country (1959 - 2024)
       x = "Country",
       y = "Total Females") +
  scale_fill_gradient(low = "lightblue", high = "darkblue", name = "Female Participat
  geom_text(aes(label = total_awards),
  position = position_dodge(width = 0.9), hjust = -0.1, color = "black", size = 3) +
  theme_minimal() +
theme(
    axis.title.x = element_blank(),
    plot.title = element_text(size = 14, face = "bold"),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    legend.position = "bottom",
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank())
```



```
country_female_participation2_filtered
```

country <chr>	total_females <dbl>	total_team_size <dbl>
Albania	35	149
Armenia	21	193
Bangladesh	13	111
Belarus	30	193
Benin	2	5
Bolivia	19	92
Bosnia and Herzegovina	33	184
Botswana	14	56
Burkina Faso	3	12
Cambodia	12	52

1-10 of 65 rows | 1-3 of 9 columns Previous **1** 2 3 4 5 6 7 Next

```
filt2 <- country_female_participation2_filtered %>%
  select(country, total_females, female_proportion, total_awards)
filt2
```

country <chr>	total_females <dbl>	female_proportion <dbl>
Albania	35	0.2348993
Armenia	21	0.1088083
Bangladesh	13	0.1171171
Belarus	30	0.1554404
Benin	2	0.4000000
Bolivia	19	0.2065217
Bosnia and Herzegovina	33	0.1793478
Botswana	14	0.2500000
Burkina Faso	3	0.2500000
Cambodia	12	0.2307692

1-10 of 65 rows | 1-3 of 4 columns Previous **1** 2 3 4 5 6 7 Next

Plot Countries with high GDI ranks but low female_proportion

The Gender Development Index (GDI) measures gender equality in human development by comparing the Human Development Index (HDI) values of women and men. Here’s a breakdown of how the GDI is interpreted based on the values you’ve provided:

Interpreting GDI Values:

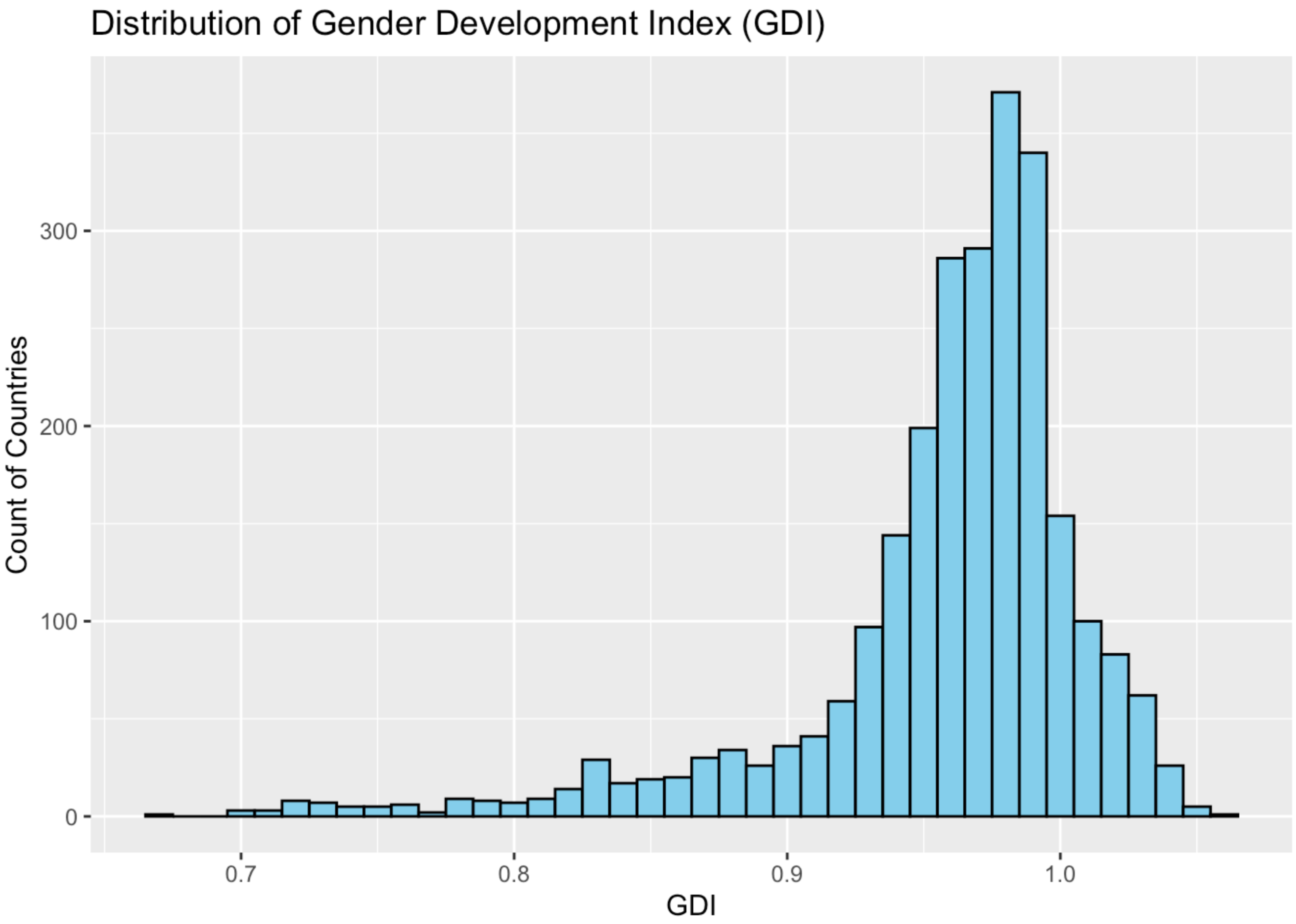
GDI < 1: Indicates that men have a higher HDI than women, suggesting a gender gap in human development favoring men. This situation implies that women may have less access to resources, opportunities, or well-being. GDI = 1: Indicates perfect gender equality, where women and men have the same HDI. GDI > 1: Indicates that women have a higher HDI than men, suggesting a gender gap in human development favoring women. This situation might reflect conditions where women experience better access to education, health care, and economic opportunities compared to men. https://ourworldindata.org/human-development-index#the-gender-development-index-gdi

```
# Create GDI categories
country_results_gdi <- country_results_gdi %>%
  mutate(gdi_category = case_when(
    gender.development.index < 1 ~ "GDI < 1: Gender Gap Favoring Men",
    gender.development.index == 1 ~ "GDI = 1: Perfect Gender Equality",
    gender.development.index > 1 ~ "GDI > 1: Gender Gap Favoring Women"
  ))

# Map countries to developing regions

# Create a column for "region"
country_results_gdi <- country_results_gdi %>%
  mutate(region = case_when(
    standardized_country %in% c("Algeria", "Bahrain", "Egypt", "Iraq", "Jordan", "Kuw
    standardized_country %in% c("Brunei", "Cambodia", "China", "Fiji", "Indonesia", '
    standardized_country %in% c("Albania", "Armenia", "Azerbaijan", "Belarus", "Bosni
    standardized_country %in% c("Antigua and Barbuda", "Argentina", "Bahamas", "Barba
    standardized_country %in% c("Afghanistan", "Bangladesh", "Bhutan", "India", "Iran
    standardized_country %in% c("Angola", "Benin", "Botswana", "Burkina Faso", "Burur
    standardized_country %in% c("Australia", "New Zealand") ~ "Oceania",
    standardized_country %in% c("United States", "Canada") ~ "North America",
    TRUE ~ "Other"  # Default if the country doesn't match any group
  ))

# View the dataframe with the new region column
head(country_results_gdi)
```

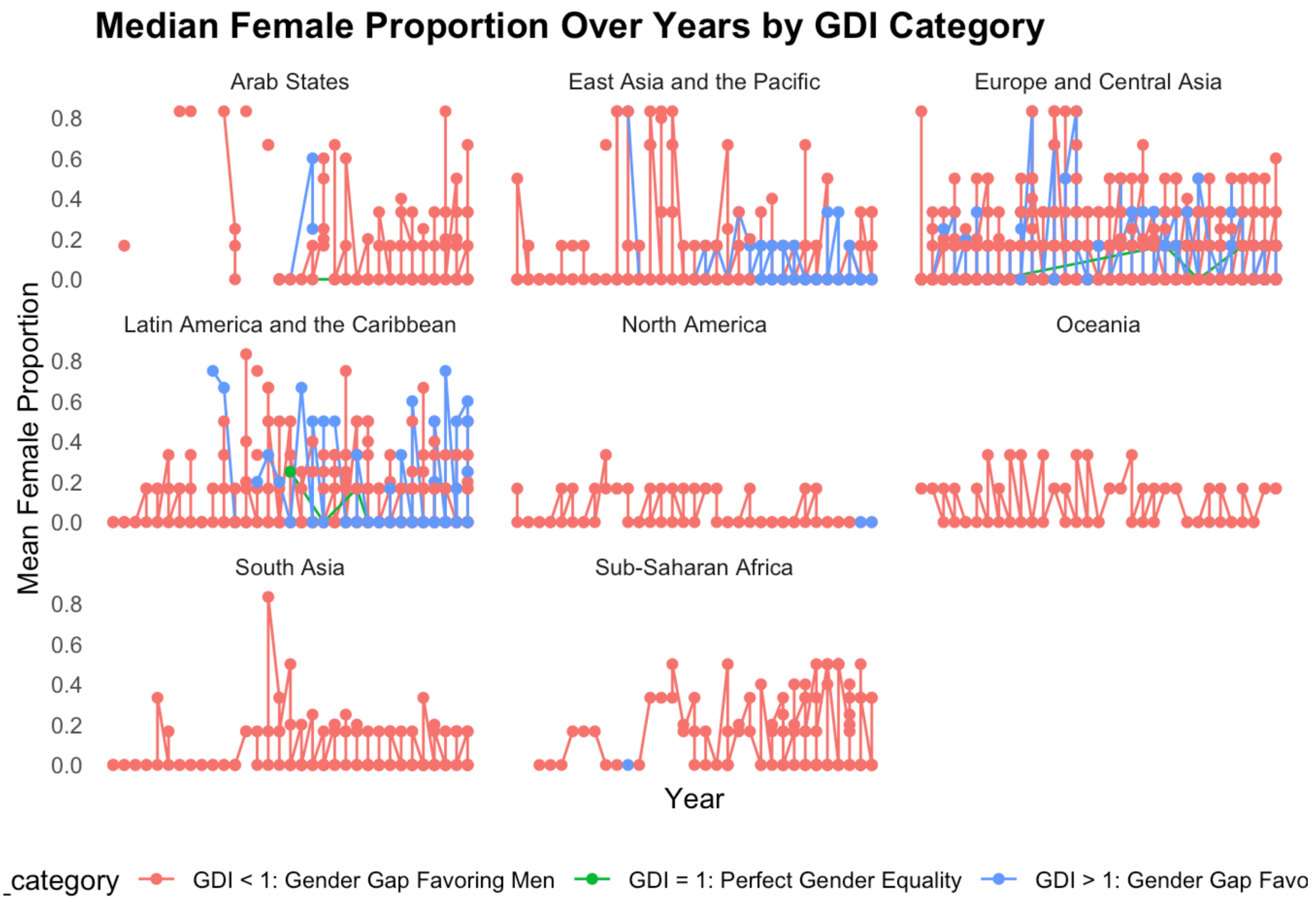
year	country	team_size_all	team_size_male
<dbl>	<chr>	<dbl>	<dbl>
2022	People's Republic of China	6	6
2022	Republic of Korea	6	6
2022	United States of America	6	6
2022	Vietnam	6	6
2022	Romania	6	6
2022	Thailand	6	6

6 rows | 1-4 of 27 columns

```
ggplot(country_results_gdi, aes(x = gender.development.index)) +
  geom_histogram(binwidth = 0.01, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Gender Development Index (GDI)",
       x = "GDI", y = "Count of Countries")
```



```
gender_proportion_yearly_summary <- country_results_gdi %>%
  group_by(year, standardized_country, gdi_category, region) %>%
  summarise(
    median_female_proportion = median(female_proportion, na.rm = TRUE),
    .groups = 'drop'  # Ungroup the data after summarization
  )
# Plot the mean female proportion over the years by GDI category
ggplot(gender_proportion_yearly_summary, aes(x = year, y = median_female_proportion,
  geom_line() +
  geom_point() +
  labs(title = "Median Female Proportion Over Years by GDI Category",
       x = "Year",
       y = "Mean Female Proportion") +
  theme_minimal() +
  facet_wrap(~ region)+
  theme(
    plot.title = element_text(size = 14, face = "bold"),
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    legend.position = "bottom",
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank())
```



```
gender_proportion_yearly_summary
```

year	standardized_country	gdi_category
<dbl>	<chr>	<chr>
1990	Argentina	GDI < 1: Gender Gap Favoring Men

year	standardized_country	gdi_category
<dbl>	<chr>	<chr>
1990	Australia	GDI < 1: Gender Gap Favoring Men
1990	Austria	GDI < 1: Gender Gap Favoring Men
1990	Bahrain	GDI < 1: Gender Gap Favoring Men
1990	Bulgaria	GDI < 1: Gender Gap Favoring Men
1990	Canada	GDI < 1: Gender Gap Favoring Men
1990	China	GDI < 1: Gender Gap Favoring Men
1990	Cyprus	GDI < 1: Gender Gap Favoring Men
1990	Finland	GDI > 1: Gender Gap Favoring Women
1990	France	GDI < 1: Gender Gap Favoring Men

1-10 of 2,557 rows | 1-3 of 5 columns

Previous **1** 2 3 4 5 6 ... 256 Next

```
other <- gender_proportion_yearly_summary %>% filter(region == "Other")
other
```

```
0 rows
```