Math Olympiad

This data is sourced from the International Mathematical Olympiad (IMO), the global championship for high school students in mathematics. Held annually in a different country, the IMO began in 1959 in Romania with participation from seven countries. Over time, it has grown to include more than 100 countries across five continents. The competition spans two consecutive days, featuring six problems in total, with three problems tackled each day. The Official website for IMO is https://www.imo-official.org/ (https://www.imo-official.org/).

More information about the dataset can be found here:
https://github.com/rfordatascience/tidytuesday/tree/main/data/2024/2024-09-24
(https://github.com/rfordatascience/tidytuesday/tree/main/data/2024/2024-09-24)

Inspiration As a math major with a deep passion for mathematics, I wanted my first R project to incorporate a mathematical theme. What better dataset to explore than the International Mathematical Olympiad (IMO)? During my school years, I also participated in Kangaroo and district-level math competitions, which further fueled my interest in this project.

Question:

1. How have country rankings shifted over time?
2. What is the distribution of participation by country and gender? What's the distribution of top scores?

Introduction: I am using three datasets related to the International Mathematical Olympiad (IMO) to address the research questions. These datasets include IMO data based on countries (country_results), individuals (individual_results), and years (timeline).

To answer the first question, "How have country rankings shifted over time?", I used the country_results dataset. The variables included are country, awards_gold, awards_silver, awards_bronze, awards_honorable_mentions and year. Country is a categorical variable representing the participating country. Awards_gold, wards_silver, awards_bronze, awards_honorable_mentions are numeric variables representing the number of gold, silver, and bronze medals, as well as honorable mentions for each country. Year is a numeric variable indicating the year of the IMO.

To answer the second question, "What is the distribution of participation by country and gender? What's the distribution of top scores?", I utilized two datasets: timeline and individual_results. For distribution of participation by country, I used the variables from the timeline dataset - year and countries. Countries is a numeric variable showing the number of countries participating each year in the IMO. For distribution of participation by gender, I used the variables from the timeline dataset - year, male_contestant and female_contestant. Male_contestant and female_contestant are numeric variables indicating the number of male and female participants each year. For distribution of top scores, I referred to the individual_results dataset, focusing on year and total. Total is a numeric variable showing the total score achieved by each individual at the IMO.

Approach: My approach began with understanding and cleaning the data to prepare it for further analysis. Based on my research, each International Mathematical Olympiad (IMO) consists of only six questions, which made the p7 column in both the country_results and individual_results datasets irrelevant. Moreover, most of its records contained NA values. Therefore, I removed this column. Additionally, I excluded data prior to 1980 as it contained invalid values, such as NA or scores greater than 7, which is impossible given the maximum score for each question is 7. In the individual_results dataset, I further removed records where p1 had NA values. To make the awards data more interpretable, I transformed the award column to display four categories: Gold Medal, Silver

Medal, Bronze Medal, and Honorable Mention. In the timeline dataset, I identified missing data for the number of female contestants in 1968 and decided to remove this record for both genders to avoid introducing bias in the gender-based analysis.

To answer the first question, “How have country rankings shifted over time?”, I began by analyzing data for the year 2024. I created a bar plot to visualize the number of gold medals won by each participating country in that year. To further explore performance trends, I selected the top four countries with the best results in 2024 and created a line chart to show their performance over time, from 2000 to 2024. This chart displayed the distribution of their awards across all four categories: Gold Medal, Silver Medal, Bronze Medal, and Honorable Mention. This approach helped reveal trends and shifts in the rankings of these top-performing countries.

For the second question, “What is the distribution of participation by country and gender? What’s the distribution of top scores?”, I generated three visualizations. First, I created a line chart to display the distribution of participating countries over the years. Next, I developed a density plot to analyze the distribution of participation by gender, which provided insights into male and female contestant trends across different years. Finally, I created a violin chart to visualize the distribution of gold medal scores over time. These visualizations helped in understanding the overall trends, participation patterns, and performance of contestants from 2000 to 2024, ensuring a consistent and accurate analysis.

Analysis:

```
# Data Cleaning
country_results <- country_results %>%
  select(-c(p7)) %>%
  filter(year >= 1980)
country_results
```

```
## # A tibble: 3,507 × 17
##     year country team_size_all team_size_male team_size_female    p1    p2    p3
##    <dbl> <chr>           <dbl>          <dbl>            <dbl> <dbl> <dbl> <dbl>
##  1  2024 United…             6              5                1    42    41    19
##  2  2024 People…             6              6                0    42    42    31
##  3  2024 Republ…             6              6                0    42    37    18
##  4  2024 India               6              6                0    42    34    11
##  5  2024 Belarus             6              6                0    42    30    10
##  6  2024 Singap…             6              6                0    42    37     7
##  7  2024 United…             6              6                0    42    33     8
##  8  2024 Hungary             6              6                0    42    37    16
##  9  2024 Poland              6              6                0    42    25     5
## 10  2024 Türkiye             6              5                1    38    37     5
## # ℹ 3,497 more rows
## # ℹ 9 more variables: p4 <dbl>, p5 <dbl>, p6 <dbl>, awards_gold <dbl>,
## #   awards_silver <dbl>, awards_bronze <dbl>, awards_honorable_mentions <dbl>,
## #   leader <chr>, deputy_leader <chr>
```

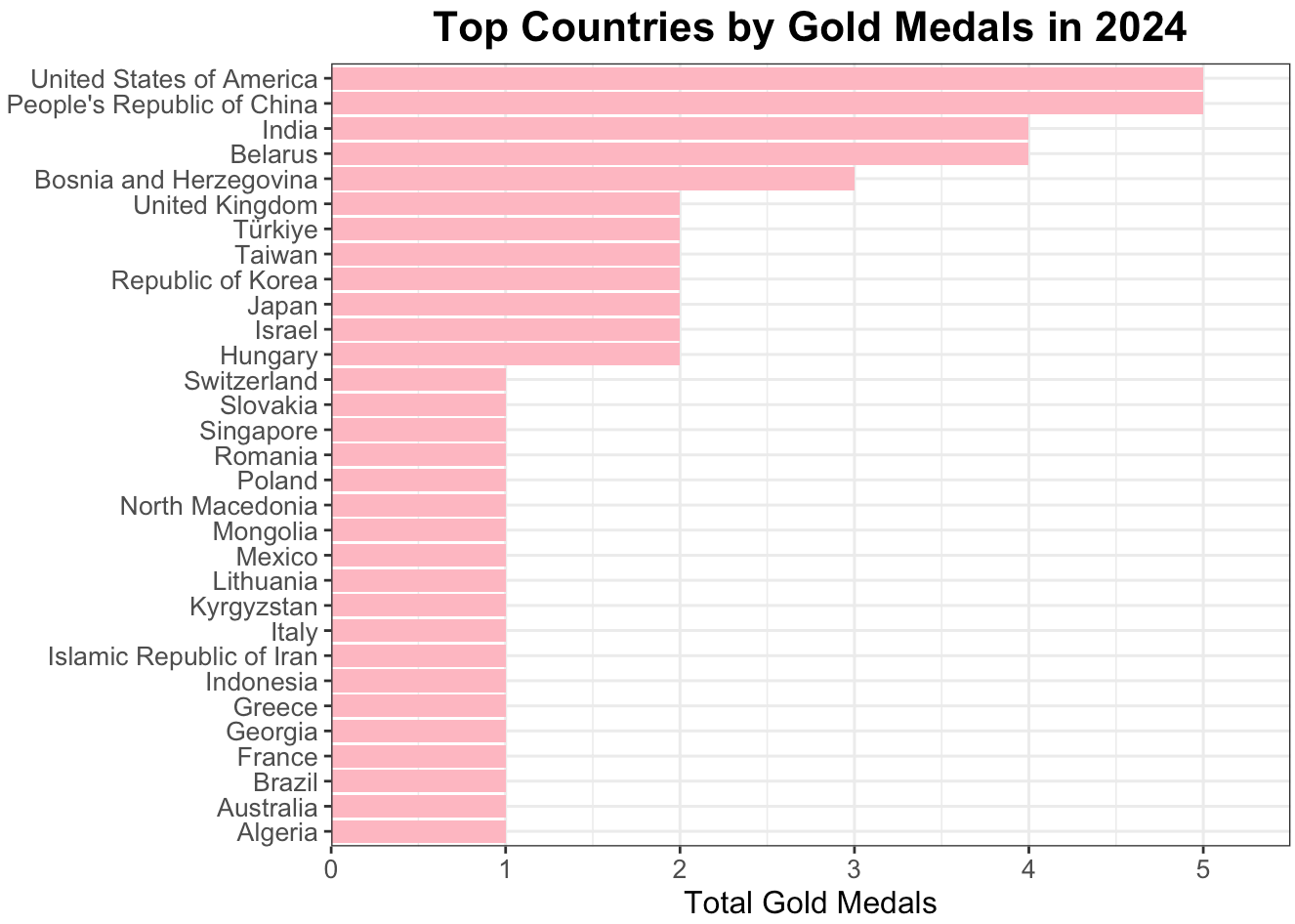
```
individual_results <- individual_results %>%
  filter(year >= 1980,
         !is.na(p1))  %>%
  select(-c(p7)) %>%
  mutate (award = case_when(
    award %in% c("Bronze medal", "Bronze medal¬ß,", "Bronze medal, Special prize"
  ) ~ "Bronze",
  award %in% c("Gold medal", "Gold medal, Special prize", "Gold medal, Special prize
(2)") ~ "Gold",
   award %in% c("Honourable mention", "Honourable mention¬ß", "Special prize") ~ "Honora
ble mention",
  award %in% c("Silver medal", "Silver medal, Special prize", "Silver medal, Special pri
ze (2)") ~ "Silver",
  TRUE ~ NA_character_)) %>%
  filter(!is.na(award))
individual_results
```

```
## # A tibble: 13,442 × 12
##     year contestant            country    p1    p2    p3    p4    p5    p6 total
##    <dbl> <chr>                 <chr>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1  2024 Haojia Shi            People…     7     7     7     7     7     7    42
##  2  2024 Ivan Chasovskikh      C21         7     7     6     6     7     7    40
##  3  2024 Alexander Wang        United…     7     7     3     7     7     7    38
##  4  2024 Satoshi Kano          Japan       7     7     2     7     7     7    37
##  5  2024 László Bence Simon    Hungary     7     7     7     7     7     0    35
##  6  2024 Adhitya Mangudy Venk… India       7     7     4     7     7     3    35
##  7  2024 Qiming Xu             People…     7     7     7     7     7     0    35
##  8  2024 Hyeongjoe Chu         Republ…     7     2     7     7     7     5    35
##  9  2024 Alex Chui             United…     7     7     2     7     7     5    35
## 10  2024 Jessica Wan           United…     7     7     5     7     7     2    35
## # ℹ 13,432 more rows
## # ℹ 2 more variables: individual_rank <dbl>, award <chr>
```

```
timeline <- timeline %>%
  filter (year != 1968)
timeline
```

```
## # A tibble: 64 × 10
##    edition  year country          city  countries all_contestant male_contestant
##      <dbl> <dbl> <chr>            <chr>     <dbl>          <dbl>            <dbl>
##  1      65  2024 United Kingdom   Bath        108            609              528
##  2      64  2023 Japan            Chiba       112            618              550
##  3      63  2022 Norway           Oslo        104            589              521
##  4      62  2021 Russian Federat… A di…       107            619              555
##  5      61  2020 Russian Federat… A di…       105            616              560
##  6      60  2019 United Kingdom   Bath        112            621              556
##  7      59  2018 Romania          Cluj…       107            594              535
##  8      58  2017 Brazil           Rio …       111            615              553
##  9      57  2016 Hong Kong        Hong…       109            602              531
## 10      56  2015 Thailand         Chia…       104            577              525
## # ℹ 54 more rows
## # ℹ 3 more variables: female_contestant <dbl>, start_date <date>,
## #   end_date <date>
```

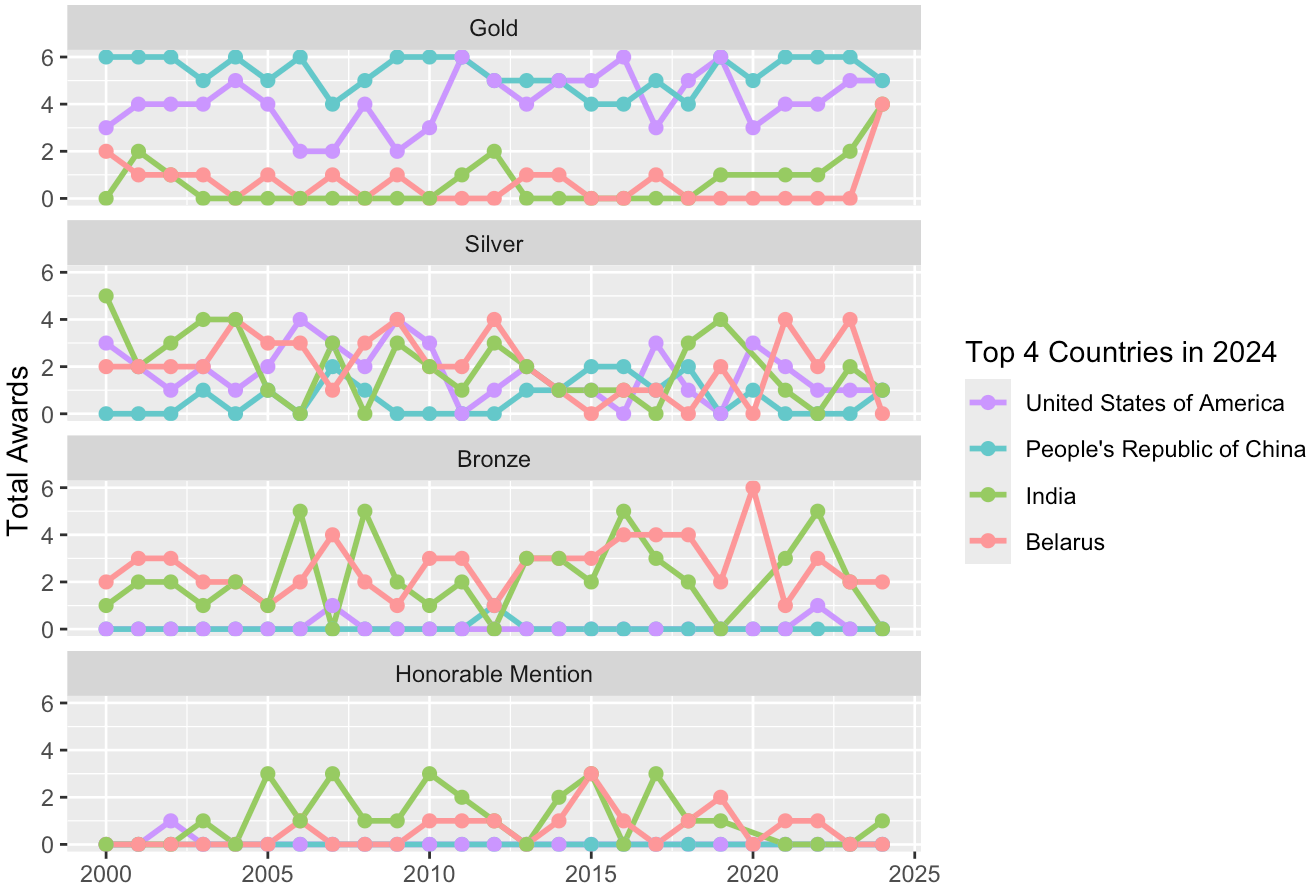
```
#Bar Char
p1 <- country_results %>%
  filter(year == "2024", awards_gold > 0) %>%
  select(country, awards_gold) %>%
  mutate(country = fct_reorder(country, awards_gold)) %>%
  ggplot(aes(x=awards_gold, y=country)) +
  geom_col(fill = "light pink") +
  ggtitle("Top Countries by Gold Medals in 2024") +
  scale_x_continuous(name = "Total Gold Medals",
                     expand = expansion(mult = c(0, 0.1))) +
  scale_y_discrete(name = NULL) +
  theme_bw() +
  theme(
    plot.title = element_text(size = 16, face = "bold", hjust = 0.5),
    axis.text.y = element_text(size = 10),
    axis.text.x = element_text(size = 10),
    axis.title = element_text(size = 12)
  )
p1
```


Top Countries by Gold Medals in 2024
United States of America
People's Republic of China
India
Belarus
Bosnia and Herzegovina
United Kingdom
Türkiye
Taiwan
Republic of Korea
Japan
Israel
Hungary
Switzerland
Slovakia
Singapore
Romania
Poland
North Macedonia
Mongolia
Mexico
Lithuania
Kyrgyzstan
Italy
Islamic Republic of Iran
Indonesia
Greece
Georgia
France
Brazil
Australia
Algeria
0
1
2
3
4
5
Total Gold Medals

```r
# Top 4 countries with gold medals in 2024
top_4_countries <- country_results %>%
  filter(year == "2024") %>%
  select(country, awards_gold) %>%
  arrange(desc(awards_gold))  %>%
  slice(1:4)

# Line Chart
p2 <- country_results %>%
  filter(
    year >= 2000,
    country %in% as.vector(top_4_countries$country))  %>%
  pivot_longer(cols = c(awards_gold, awards_silver, awards_bronze, awards_honorable_ment
ions),
               names_to = "awards",
               values_to = "award_count") %>%
  select (year, country, awards, award_count) %>%
  mutate(awards = factor(awards,
                         levels =  c("awards_gold", "awards_silver", "awards_bronze", "a
wards_honorable_mentions"),
                         labels = c("Gold", "Silver", "Bronze", "Honorable Mention"))) %
>%
  mutate(country = factor(country,
                          levels =  c("United States of America", "People's Republic of
China", "India", "Belarus"))) %>%
  ggplot(aes(x = year, y= award_count, color = country, group = country)) +
  geom_line(linewidth = 1) +
  geom_point(size = 2) +
  facet_wrap(~ awards, nrow = 4) +
  ggtitle("Top Performers of 2024: Yearly Awards Distribution") +
  scale_x_continuous(name = NULL,
                     limits = c(2000, 2024)) +
  scale_y_continuous(name = "Total Awards",
                     limits = c(0,6),
                     breaks = c(0,2,4,6)) +
  scale_color_manual(name = "Top 4 Countries in 2024",
                     values = c("Belarus" = "#FF9999",
                                "India" = "#99CC66",
                                "People's Republic of China" = "#66CCCC",
                                "United States of America" = "#CC99FF")) +
  theme(
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5)
  )
p2
```

Top Performers of 2024: Yearly Awards Distribution



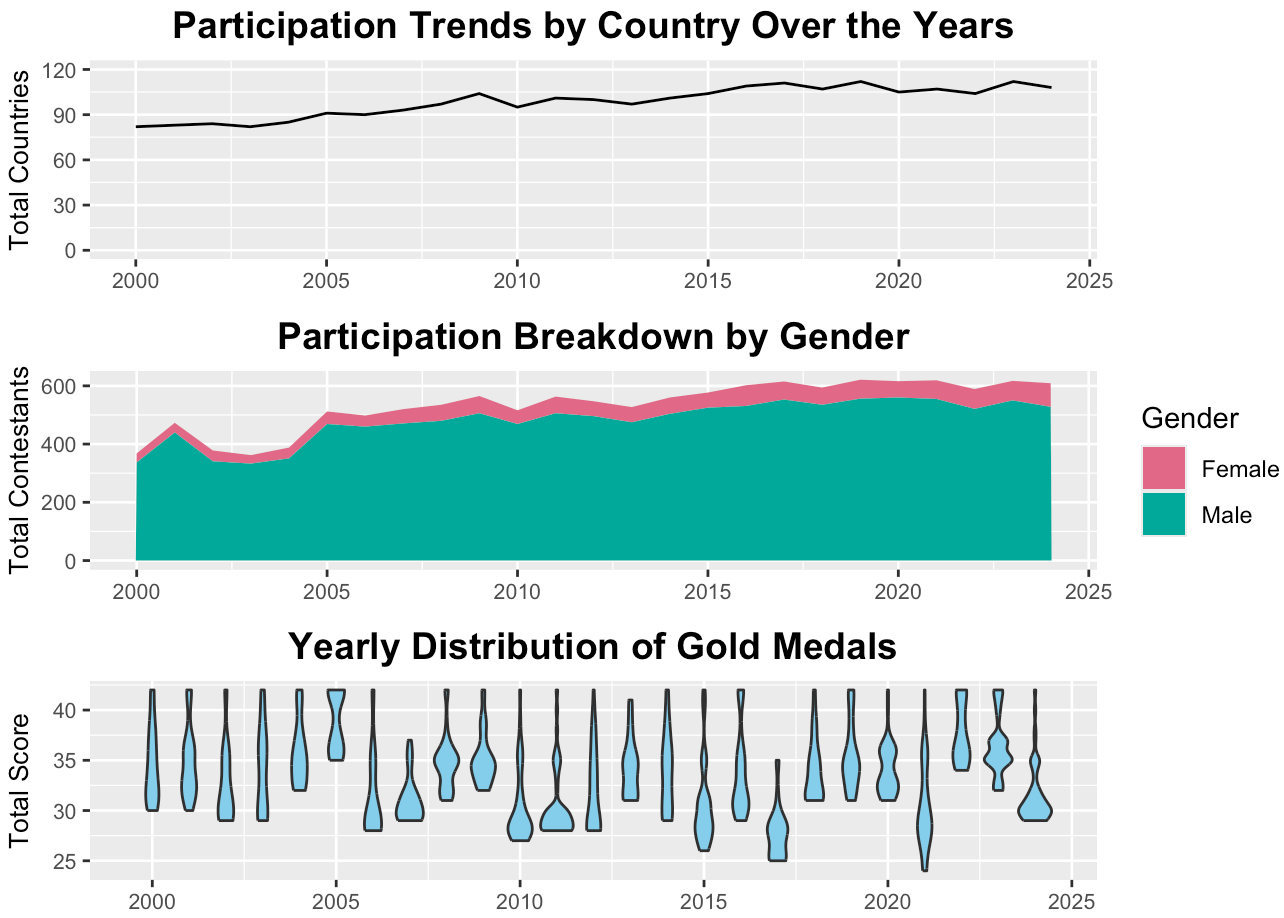
```r
# Line Chart
p3 <- timeline %>%
  filter(year >=2000) %>%
  ggplot(aes(year, countries)) +
  geom_line() +
  ggtitle("Participation Trends by Country Over the Years") +
  scale_y_continuous(name = "Total Countries",
                     limits = c(0, 120),
                     breaks = c(0, 30, 60, 90, 120)) +
  scale_x_continuous(name = NULL) +
  theme(
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5),
    axis.text.y = element_text(size = 8),
    axis.text.x = element_text(size = 8),
    axis.title = element_text(size = 10)
  )

# Density Plot
p4 <- timeline %>%
  filter(year >=2000) %>%
  pivot_longer(cols = c(male_contestant, female_contestant),
               names_to = "Gender",
               values_to = "Contestant_count") %>%
  select (year, country, Gender, Contestant_count) %>%
  ggplot(aes(year,Contestant_count, fill = Gender)) +
  geom_area() +
  ggtitle("Participation Breakdown by Gender") +
  scale_y_continuous(name = "Total Contestants") +
  scale_x_continuous(name = NULL) +
  scale_fill_discrete_qualitative(name = "Gender", labels = c("Female", "Male")) +
  theme(
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5),
    axis.text.y = element_text(size = 8),
    axis.text.x = element_text(size = 8),
    axis.title = element_text(size = 10)
  )

# Violin
p5 <- individual_results %>%
  filter(year >=2000,
         award == "Gold") %>%
  select(year, total) %>%
  ggplot(aes(x=year, y=total, group = year)) +
  geom_violin(fill = "skyblue") +
  scale_y_continuous(name = "Total Score") +
  scale_x_continuous(name = NULL) +
  ggtitle("Yearly Distribution of Gold Medals") +
  theme(
    plot.title = element_text(size = 14, face = "bold", hjust = 0.5),
    axis.text.y = element_text(size = 8),
    axis.text.x = element_text(size = 8),
    axis.title = element_text(size = 10)
```

```
  )

p3 / p4 / p5
```



Discussion: Distribution of Gold Medals for 2024: The bar chart illustrates the total number of gold medals won by different countries in 2024. The United States and China lead with 5 gold medals each, followed closely by India and Belarus, which secured 4 gold medals each. These four countries emerged as the top performers in 2024, with the chart displaying the remaining countries in descending order of gold medal counts.

Top Performers of 2024: Yearly Awards Distribution: This visualization presents the distribution of gold, silver, bronze, and honorable mention awards for the top four countries of 2024 from 2000 to 2025. The United States consistently leads in gold medals but shows fluctuations, with peaks and troughs over time. China demonstrates strong competition with steady and consistent performances. India and Belarus, while typically earning 0–2 gold medals, stand out in 2024 with improved results. For silver medals, China displays a stronger and more consistent performance compared to the other countries, while the United States shows occasional dips, particularly during the mid-2010s. Belarus has an interesting trend, often securing more silver medals relative to gold. India exhibits a fluctuating pattern for silver medals. In the bronze medal category, Belarus shows occasional spikes, whereas India struggles with maintaining consistent performance. The honorable mention category reveals irregular trends, particularly for India and Belarus, while the United States and China have historically focused more on gold and silver achievements. Overall, this visualization highlights how rankings and award distributions among these leading countries have shifted over time, offering insight into the competitive landscape of the IMO.

Participation Trends by Country Over the Years: The line graph depicts the number of countries participating in the IMO from 2000 to 2025. Participation trends remain relatively stable, with a gradual increase from around 80 countries in the early 2000s to approximately 100 in the 2020s. While minor fluctuations occur, overall

participation levels have been maintained over the years.

Participation Breakdown by Gender: A stacked area chart displays the breakdown of IMO participants by gender over time. Male contestants consistently dominate participation numbers, while female contestants form a smaller, yet growing proportion. In recent years, there has been a slight increase in female representation. The total number of contestants saw a gradual rise post-2000 but plateaued after 2015, stabilizing around 600 participants annually.

Yearly Distribution of Gold Medals: The violin plot shows the distribution of total scores for gold medalists over time. While gold medal scores generally cluster between 30 and 40, the range of scores varies year-to-year. Some years, such as 2005 and 2010, show narrower distributions, indicating less variability in performances. In contrast, years with wider violins suggest a broader spread of scores among gold medalists. Despite these variations, the trends in gold medal scores remain relatively consistent, with most scores falling within the 30–40 range.