Human Trafficking: Who's at Stake and Who's Next?

Gullit Navarrete

5/16/25

Abstract

Across the world and still lingering into the modern age, Human trafficking is a pervasive crime that ensnares people of every age, gender, and background in forced labor, commercial sex, and other formats alike to modern slavery. Today, it's estimated that 27.6 million individuals suffer under trafficking conditions. Perpetrators aren't limited to organized criminal networks, in fact some world powers and governments secretly obtain these victims including children into exploitative labor or other illicit activities. Given the global scale and severity of this injustice, it demands urgent attention and collective action to stem its spread and protect vulnerable populations.

UNODC
United Nations Office on Drugs and Crime

Introduction

- The first data source, a .CSV file, was downloaded from the United Nations Office on Drugs and Crime (UNODC) website: https://dataunodc.un.org/dp-trafficking-persons This data was collected by using the Questionnaire for the Global Report on Trafficking in Persons (GLOTIP), hence the name of the original file being "data_glotip".
- The second data source, via web scraping and using the initial source as an API, was found from the US Human Trafficking Hotline: https://humantraffickinghotline.org/en/statistics
- Dependent: Amount of trafficking victims
- Independent variables: Gender (Sex), Age, and year for linear regression.

Data Science Workflow:

Data Import:

- The .CSV was downloaded from the United Nations Office on Drugs and Crime (UNDOC) website and read into R.
- The API for the US Human Trafficking Hotline is imported with a rvest approach, because the website/source doesn't expose those year-by-year numbers via a neat JSON endpoint. So therefore, I scraped out the data using rvest by reading the website into R, then grabbing each year's panel and extracting.

Data Transformation:

- After importing both the UNODC CSV and the scraped Hotline data, I tidied each dataset into a common "long" format and evened out both their key fields. For the UNODC table, I filtered to only U.S. rows, "Detected trafficking victims," and total Sex/Age dimensions, then renamed columns for clarity. For the Hotline data, I will first removed the redundant "Unit of measurement" column (all entries were "Counts" anyways), renamed Iso3_code to Abbreviation (which is basically a 3-4 letter code for a country) and txtVALUE to Count, and converted "<5" into a numeric midpoint value of 2 before parsing each text string to a number. I will then removed the placeholder "NA" panel (the 2007–2014 aggregate) and filter out any remaining NA years, producing two clean, year based time series data.

Data Analysis:

- The research questions; 1. What is the difference in counts between the UNODC vs. the US Human Trafficking Hotline 2. What are some trends over time for gender and/or age? 3. What are the results of t-tests that compare males to females and ages 17 and less to ages 18 and older? would be answered using summary statistics as well as visualizing the data to answer these questions. For the final research question: "How can you use a linear regression line to determine an average number of people in human trafficking (across all countries and in the United States) for each of the following: Males 17 or younger, males 18 or older, females 17 or younger, and females 18 or older" I'll be using linear regression as a analytical method to predict.

Data Import

Data Import

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	UNODC	unodc_da	ss@un.org										
2	14/04/2025												
3	Iso3_code	Country	Region	Subregion	Indicator	Dimensio	Category	Sex	Age	Year	Unit of me	txtVALUE	Source
4	ABW	Aruba	Americas	Latin Ame	Detected t	by country	Ukraine	Total	Total	2010	Counts	<5	GLOTIP
5	AFG	Afghanista	Asia	Southern A	Detected t	by country	Abroad	Total	Total	2003	Counts	<5	GLOTIP
6	AFG	Afghanista	Asia	Southern A	Detected t	by country	Abroad	Total	Total	2008	Counts	<5	GLOTIP
7	AFG	Afghanista	Asia	Southern A	Offences c	Total	Total	Total	Total	2013	Counts	103	GLOTIP
8	AFG	Afghanista	Asia	Southern A	Persons b	Total	Total	Total	Total	2013	Counts	167	GLOTIP
9	AFG	Afghanista	Asia	Southern A	Detected t	by form of	Sexual exp	Total	Total	2017	Counts	<5	GLOTIP
10	AFG	Afghanista	Asia	Southern A	Detected t	by form of	Forced lat	Total	Total	2018	Counts	<5	GLOTIP
11	AFG	Afghanista	Asia	Southern A	Detected t	by form of	Sexual exp	Total	Total	2018	Counts	<5	GLOTIP
12	AFG	Afghanista	Asia	Southern A	Detected t	by form of	Sexual exp	Total	Total	2019	Counts	<5	GLOTIP
13	AFG	Afghanista	Asia	Southern A	Detected t	by form of	Forced lat	Total	Total	2020	Counts	<5	GLOTIP
14	AGO	Angola	Africa	Sub-Sahar	Detected t	by country	Abroad	Total	Total	2003	Counts	<5	GLOTIP
15	AGO	Angola	Africa	Sub-Sahar	Detected t	by country	Abroad	Total	Total	2008	Counts	<5	GLOTIP
16	AGO	Angola	Africa	Sub-Sahar	Detected t	Total	Total	Total	0 to 17 yea	2009	Counts	15	GLOTIP
17	AGO	Angola	Africa	Sub-Sahar	Detected t	Total	Total	Total	18 years o	2009	Counts	<5	GLOTIP

```{r load-data}
github <- "https://raw.githubusercontent.com/GullitNa/DATA607FINAL/main/data_glotip.csv"
data_glotip <- read.csv(
  github,
  skip = 2,
  stringsAsFactors = FALSE,
  check.names = FALSE
)
head(data_glotip)
```

Description: df [6 × 13]

	Iso3_code <chr>	Country <chr>	Region <chr>	Subregion <chr>
1	ABW	Aruba	Americas	Latin America and the Caribbean
2	AFG	Afghanistan	Asia	Southern Asia
3	AFG	Afghanistan	Asia	Southern Asia
4	AFG	Afghanistan	Asia	Southern Asia
5	AFG	Afghanistan	Asia	Southern Asia
6	AFG	Afghanistan	Asia	Southern Asia

6 rows | 1-5 of 13 columns

```{r}
unique(data_glotip$`Unit of measurement`)
```

```
[1] "Counts"
```

Only has Counts so I'll remove this column in Data Transformation.

Data Import

```r
url    <- "https://humantraffickinghotline.org/en/statistics"
page   <- read_html(url)
panels <- page %>% html_nodes("section.js-tabs-panel")

ids    <- panels %>% html_attr("id")
years <- as.integer(str_remove(ids, "^year\\-"))
counts <- lapply(panels, function(p) {
  p %>%
    html_nodes(".text-h1.font-black") %>%
    html_text(trim = TRUE) %>%
    str_remove_all(",") %>%
    as.integer()
})

yearly_stats <- data.frame(
  Year              = years,
  Signals_Received  = sapply(counts, `[`, 1),
  Victim_Signals    = sapply(counts, `[`, 2),
  Cases_Identified  = sapply(counts, `[`, 3),
  Victims           = sapply(counts, `[`, 4),
  row.names = NULL
)
yearly_stats
```

Description: df [10 × 5]

Year <int>	Signals_Received <int>	Victim_Signals <int>	Cases_Identified <int>	Victims <int>
2023	30162	7380	9619	16999
2022	37331	9647	9014	15299
2021	47325	12846	10353	16708
2020	51611	13946	10525	16984
2019	49669	10676	11368	22163
2018	41736	8172	10703	21714
2017	34017	5485	8596	21305

Remove the "NA" year (2007-2014)

```r
yearly_stats <- yearly_stats %>%
  filter(!is.na(Year))
```

- First challenge of the project for me.
- No formal API/easily accessible JSON.
- Inconsistent HTML structure & formatting (bad column reading for years and numbers treated as text that was separated by commas.
- Used rvest package to pull the full page HTML, grabbed each panel's id (to derive the numeric year) and its four .text-h1.font-black values, and cleaned them up with str_remove_all() and as.integer()

Data Transformation

Data Transformation

```
or the .CSV file, I'll begin by checking if the "Unit of measurement" column confirms my suspicion of
nly having "Counts". If so, then I'll remove it as it is a repetitive column.
``{r}
nique(data_glotip$`Unit of measurement`)
```

```
[1] "Counts"
```

```
onfirmed that the column literally only has "Counts".
``{r}
lotip <- data_glotip %>%
 rename(
   Abbreviation = Iso3_code,  # rename country-code
   Count        = txtVALUE    # rename the value column
 ) %>%
 select(
   -`Unit of measurement`     # drop the one-only "Counts" column
 ) %>%
 mutate(
   Count = parse_number(Count, na = "<5")  # turn "<5" into NA, "1,234" → 1234
 )

 peek
ead(glotip)
```

Description: df [6 × 12]

	Abbreviation <chr>	Country <chr>	Region <chr>	Subregion <chr>
1	ABW	Aruba	Americas	Latin America and the Caribbean
2	AFG	Afghanistan	Asia	Southern Asia
3	AFG	Afghanistan	Asia	Southern Asia
4	AFG	Afghanistan	Asia	Southern Asia
5	AFG	Afghanistan	Asia	Southern Asia
6	AFG	Afghanistan	Asia	Southern Asia

```
Change the "NA" values in Counts to be the midpoint (2) on average instead of just throwing away the
data entirely in analysis
```{r}
glotip <- glotip %>%
  mutate(
    Count = replace_na(Count, 2)
  )
```
```

```{r}
us_victims <- glotip %>%
  filter(
    Country   == "United States of America",
    Indicator == "Detected trafficking victims",
    Dimension == "Total",
    Sex       == "Total",
    Age       == "Total"
  ) %>%
  group_by(Year) %>%
  summarize(
    UNODC_Victims_US = sum(Count),
    .groups = "drop"
  )

# 3) Clean up the scraped hotline stats and drop that NA-year row
yearly_stats_clean <- yearly_stats %>%
  filter(!is.na(Year)) %>%
  rename(Hotline_Victims = Victims)

# 4) Stitch them together on Year
combined <- yearly_stats_clean %>%
  inner_join(us_victims, by = "Year")

# peek
combined
```

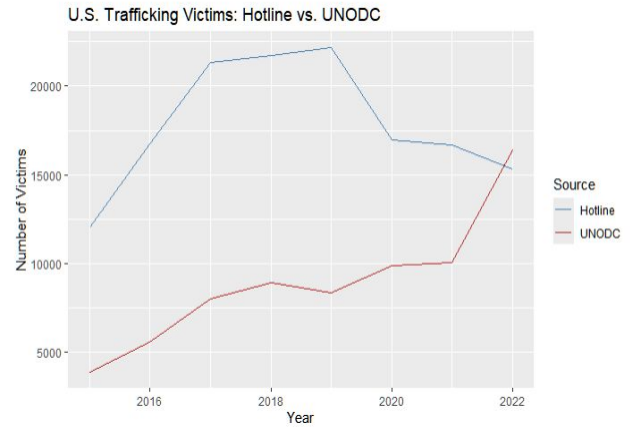
Description: df [8 × 6]

Year <int>	Signals_Received <int>	Victim_Signals <int>	Cases_Identified <int>	Hotline_Victims <int>	UNODC_Victims_US <dbl>
2022	37331	9647	9014	15299	16390
2021	47325	12846	10353	16708	10070
2020	51611	13946	10525	16984	9854
2019	49669	10676	11368	22163	8375
2018	41736	8172	10703	21714	8913
2017	34017	5485	8596	21305	8003
2016	31986	5148	7528	16747	5582
2015	26840	4469	5551	12008	3885

Data Analysis

Data Analysis: Hotline vs UNODC

```
## Hotline vs. UNODC every year
Now you truly have two datasets in one analysis—compare the hotline's identified victims against the
UNODC's official counts:
```{r}
ggplot(combined, aes(x = Year)) +
  geom_line(aes(y = Hotline_Victims, color = "Hotline")) +
  geom_line(aes(y = UNODC_Victims_US, color = "UNODC")) +
  labs(
    y = "Number of Victims",
    color = "Source",
    title = "U.S. Trafficking Victims: Hotline vs. UNODC"
  ) +
  scale_color_manual(values = c("Hotline" = "steelblue", "UNODC" = "firebrick"))
```
```



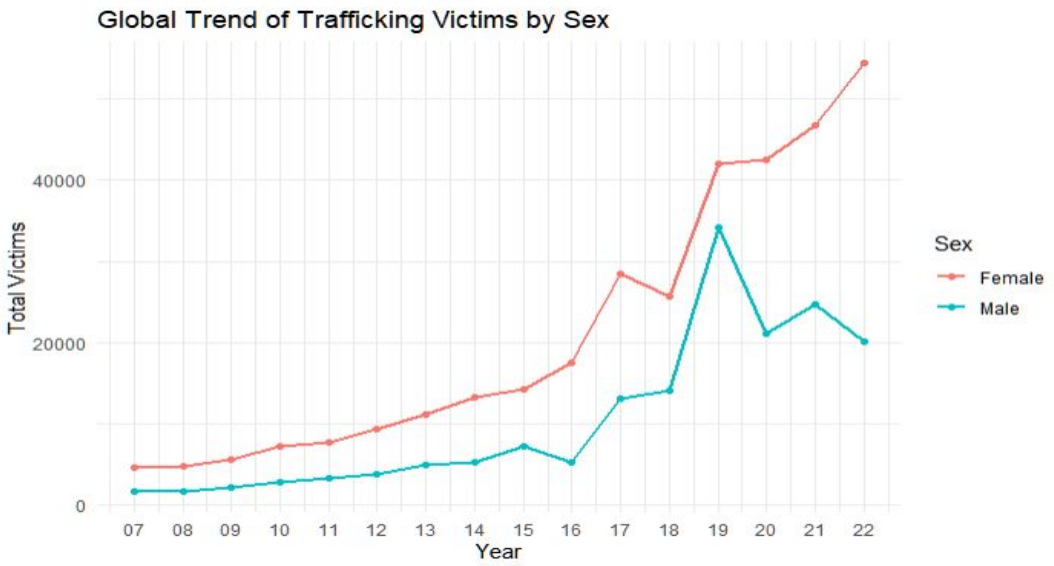
- The Hotline's total victims curve sits well above the UNODC's. This is possible because the Hotline is picking up substantially more "signals" or potential victim reports than what shows up in the UNODC's official national totals—often by a factor of two or more.
- UNODC data comes from country reports, law enforcement and government agencies, aggregated and vetted before publication. There may be "definition" differences between different countries and cultures.
- Hotline data is "real-time" outreach: it includes any contact (calls, texts, chats) that meet the trafficking-victim criteria, even if those never end up as formally "identified" cases in government statistics.
- **What does this counting gap mean?** Under-reporting in official channels: many victims never come to the attention of law enforcement but do make it to the Hotline.

Sex <chr>	Total_Victims <dbl>
Female	335046
Male	165173
Other	2382
Total	565985

4 rows

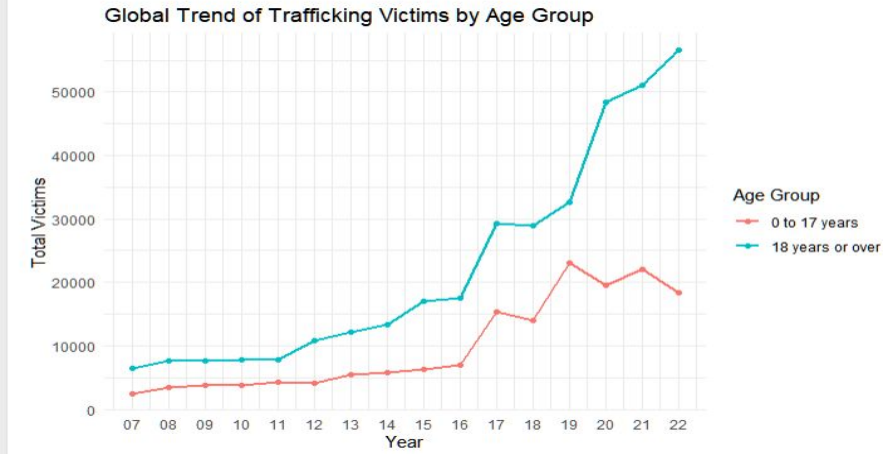
Data Analysis: Trends Sex/Gender

```{r}
df_totals <- glotip %>%
  filter(
    Indicator == "Detected trafficking victims",
    Dimension == "Total"
  )
gender_trend <- df_totals %>%
  filter(Age == "Total") %>%
  group_by(Year, Sex) %>%
  summarize(Victims = sum(Count, na.rm = TRUE), .groups = "drop")

# FOR PLOTTING ONLY
gender_plot_df <- gender_trend %>%
  filter(Sex %in% c("Female", "Male"))

ggplot(gender_plot_df, aes(x = Year, y = Victims, color = Sex)) +
  geom_line(size = 1) +
  geom_point() +
  scale_x_continuous(
    breaks = unique(gender_plot_df$Year),
    labels = function(x) sprintf("%02d", x %% 100) #better year clarity
  ) +
  labs(
    title = "Global Trend of Trafficking Victims by Sex",
    x     = "Year",
    y     = "Total Victims"
  ) +
  theme_minimal()
```



Data Analysis: Trends Age Group

```
ggplot(age_plot_df, aes(x = Year, y = Victims, color = Age)) +
  geom_line(size = 1) +
  geom_point() +
  scale_x_continuous(
    breaks = unique(age_plot_df$Year),
    labels = function(x) sprintf("%02d", x %% 100)
  ) +
  labs(
    title = "Global Trend of Trafficking Victims by Age Group",
    x     = "Year",
    y     = "Total Victims",
    color = "Age Group"
  ) +
  theme_minimal()
```

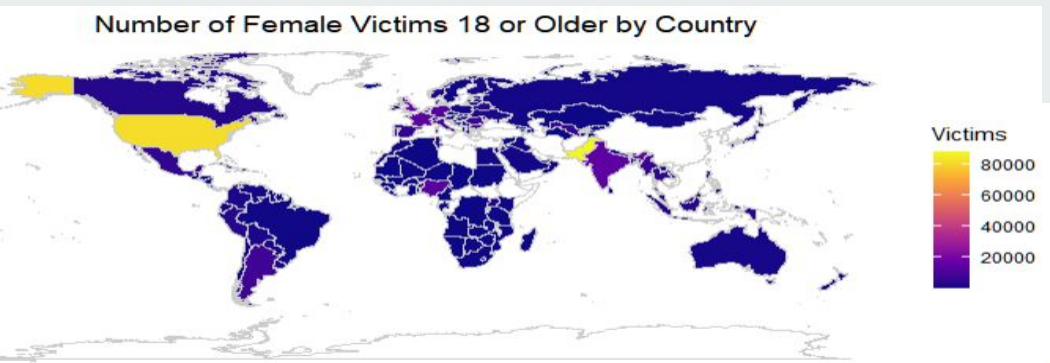


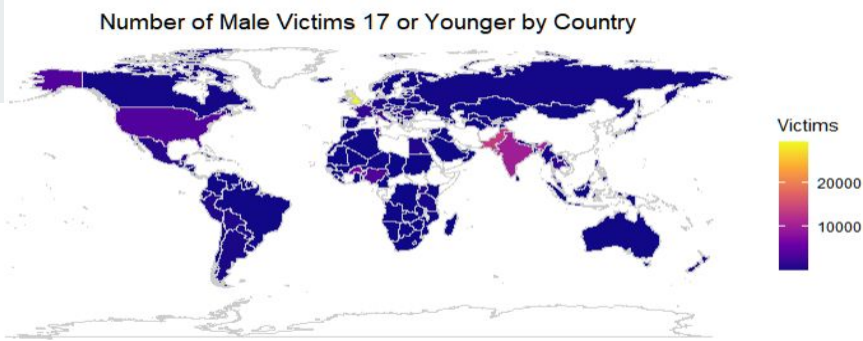
- According to the graph and summary statistics, the victims of human trafficking tend to come from the age group "18 or over" more so than those who come from the age group "0 to 17 years" or 17 or younger.

A tibble: 3 × 2

Age <chr>	Total_Victims <dbl>
0 to 17 years	158647
18 years or over	355004
Total	565985

3 rows

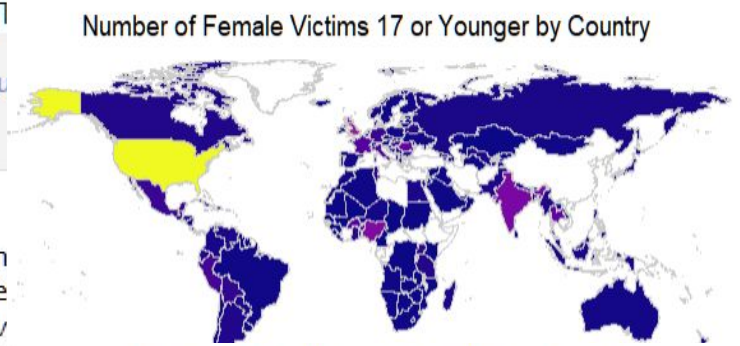


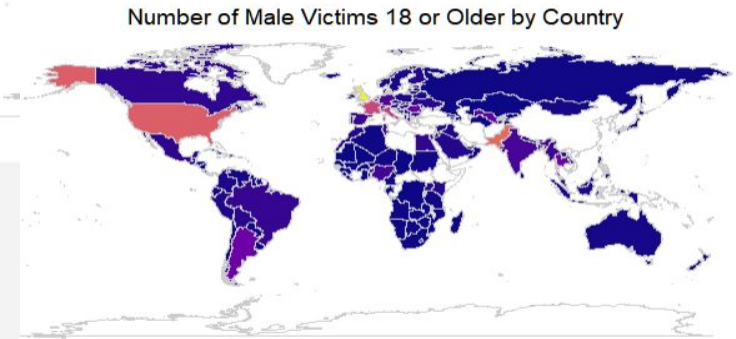


Check what is mismatched by the maps_data country names/spell

```{r}
unmatched <- anti_join(country_totals, world_outline, by="Cou
unique(unmatched$Country)
```

```
  [1] "Antigua and Barbuda"                "Hong Kong"
  [3] "Macau"                              "Republic of the Con
  [5] "Cura\xe7ao"                         "C\xf4te d\x92Ivoire
  [7] "Vatican City"                       "Saint Kitts and Nev
  [9] "Saint Vincent and the Grenadines" "Sark"
 [11] "Trinidad and Tobago"                "Tuvalu"
 [13] "T\xfcrkiye"                         "Viet Nam"
```





Manually recode

```{r}
country_totals_clean <- country_totals %>%
  mutate(
    Country = recode(Country,
      `Antigua and Barbuda` = "Antigua and Barbuda",
      `Bolivia (Plurinational State of)` = "Bolivia",
```

T-tests

- First t-test will be on the topic of these hypothesis:
- Null Hypothesis: “Males and females are victims of human trafficking in equal numbers.”
- Alternative Hypothesis: “Males and females are victims of human trafficking in different amounts.”

```
        Paired t-test

data:  sex_totals$Male and sex_totals$Female
t = -4.3628, df = 1433, p-value = 1.376e-05
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -137.62305  -52.25142
sample estimates:
mean difference
      -94.93724
```

- Null hypothesis is rejected because its p-value is < 0.001.
- The mean paired difference (Male and Female) is −94.94 (95 % CI [−137.6, −52.3]) on average there are about 95 more female victims than male victims per country-year

T-tests

- The second t-test will be on the topic of these hypothesis:
- Null Hypothesis: “Age groups 17 or younger and 18 or older are victims of human trafficking in equal numbers.”
- Alternative Hypothesis: “Age groups 17 or younger and 18 or older are victims of human trafficking in different amounts.”

```
        Paired t-test

data:  age_totals$`0 to 17 years` and age_totals$`18 years or over`
t = -6.8074, df = 1674, p-value = 1.38e-11
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -204.3603 -112.9388
sample estimates:
mean difference
      -158.6496
```

- Null hypothesis is rejected because its p-value is < 0.001.
- The mean paired difference (0 to 17 years and 18 years or over) is −158.65 (95 % CI [−204.4, −112.9]) on average there are roughly 159 more adult victims than children, per country-year.

Linear Regression

- Linear regression is a machine learning tool used for predictions, so it was an obvious choice to use this tool when it came to the research question of predicting the following scenarios: 1. Males 17 years or less in the year 2025. 2. Males 18 years or more in the year 2023. 3. Females 17 years or less in the year 2025. 4. Females 18 years or more in the year 2023.

```
global_trend <- glotip %>%
  filter(Dimension == "Total",
         Sex %in% c("Male","Female"),
         Age %in% c("0 to 17 years","18 years or over")) %>%
  group_by(Sex, Age, Year) %>%
  summarize(Victims = sum(Count, na.rm = TRUE), .groups = "drop")
mod_m_u17 <- lm(Victims ~ Year,
                data = filter(global_trend, Sex=="Male",   Age=="0 to 17 years"))
mod_m_18p <- lm(Victims ~ Year,
                data = filter(global_trend, Sex=="Male",   Age=="18 years or over"))
mod_f_u17 <- lm(Victims ~ Year,
                data = filter(global_trend, Sex=="Female", Age=="0 to 17 years"))
mod_f_18p <- lm(Victims ~ Year,
                data = filter(global_trend, Sex=="Female", Age=="18 years or over"))



preds <- tibble(
  Sex    = c("Male",   "Male",     "Female",   "Female"),
  Age    = c("0 to 17 years", "18 years or over",
             "0 to 17 years", "18 years or over"),
  Year   = c(2025,     2023,       2025,       2023),
  Predicted_Victims = c(
    predict(mod_m_u17, newdata = data.frame(Year=2025)),
    predict(mod_m_18p, newdata = data.frame(Year=2023)),
    predict(mod_f_u17, newdata = data.frame(Year=2025)),
    predict(mod_f_18p, newdata = data.frame(Year=2023))
  )
)
preds
```

Linear Regression

Sex <chr>	Age <chr>	Year <dbl>	Predicted_Victims <dbl>
Male	0 to 17 years	2025	12638.37
Male	18 years or over	2023	27219.55
Female	0 to 17 years	2025	13345.21
Female	18 years or over	2023	43781.30

4 rows

- **Male, 0–17, in 2025 = estimated 12,638 victims:** According to the model, if current year-to-year trends continue, there will be about 12.6 thousand male victims under 18 in 2025.
- **Male, 18+ in 2023 = estimated 27,220 victims:** For adult males in 2023, the model predicts 27.2 thousand victims, which shows the growing trend especially amongst those 18 or over as the victims of human trafficking when compared to 17 or under.
- **Female, 0–17 in 2025 = estimated 13,345 victims:** Predicted count for female victims under 18 in 2025 is slightly higher than the male counterpart of the same age at 13.3 thousand.
- **Female, 18+ in 2023 = estimated 43,781 victims:** Finally the model predicts that adult females in 2023 are projected at about 43.8 thousand victims significantly higher than adult males. This major disparity highlights this gender and age group's vulnerability to human trafficking, especially when compared to the 3 other subgroups.

Conclusion

Conclusion

Throughout this project, I first imported the UNODC global trafficking dataset directly from their .CSV download and then struggled but overcame with the Human Trafficking Hotline data by writing a custom rvest web scrape (since it wasn't available as a clean API endpoint). Once both sources were in R, I cleaned and reshaped the tables by renaming columns, handling "<5" counts by imputing the midpoint, and dropping redundant columns/data so that I had counts by country, year, sex/gender, and age group. I then visualized global trends with line plots (by gender and by age), ran paired t-tests to confirm that females and adults consistently experience higher victim counts, and finally built four complex maps showing male and female victims 17 or under and 18+ across every country. The closing takeaway from this project is that women aged 18 or over are by far the most vulnerable group to human trafficking worldwide, especially when compared to the other 3 subgroups within the dataset (males 17 or younger, males 18 or over, and females but 17 or younger). Of course, this doesn't exclude anyone from being a victim to human trafficking.

References

- Dataset 1 (.CSV): https://dataunodc.un.org/dp-trafficking-persons
- Dataset 2 (API/JSON): https://humantraffickinghotline.org/en/statistics
- What I used to learn about the ggplot2 map package using count by country:

  https://stackoverflow.com/questions/71858134/create-ggplot2-map-in-r-using-count-by-country