Elizabeth Harrison

Lisa Jemmings

Shirley Limburg

Saron Raut

Mara Schemel

Sumit Patel

Project 1 Write Up - Human Trafficking

For our Data Analytics Boot Camp project one, we were tasked with choosing our own set of data, cleaning the data, and providing an analysis of the data. As a group we selected our primary data source to be a data set from Kaggle provided by Counter-Trafficking Data Collaborative (CTDC). We then incorporated some additional data for country codes, GDP for countries, country governance, and population density.

Our goal was to answer the following questions:

1. How do people become trafficked? And what keeps them trapped?
2. What are the source and destination countries for human trafficking?
3. In which industries are trafficked persons most likely to be exploited?
4. Do economic factors influence the likelihood of being trafficked?

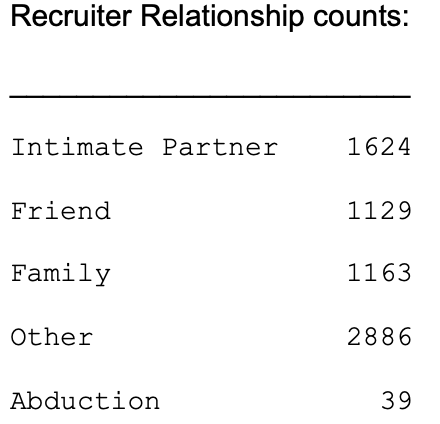
Once we determined our data set and our research questions, we needed to further explore, clean, and add to our data set. During our exploration we determined that the count of cases reported over the years were inconsistent and would be unreliable for analyzing trends over years. Therefore we decided to focus on trends across age, gender, and geography. And how these data elements related to industry, methods of control, recruitment relationship, and industry. Additionally, based on the number of cases provided from 2019 compared to the number of cases provided from the years 2002-2018, we decided to exclude 2019 cases.

The data set selected used -99 (both as an integer and as a string) to indicate missing data points. To make these values easily used during data analysis these values were replaced. When -99 was used as a string the value was replaced with ‘unknown’. This was chosen based on the columns where -99 appeared as a string. When -99 was used as an integer, it was replaced with 0. The last overall change made was to clean some of the columns containing concatenated values to make the analysis of the values more meaningful.

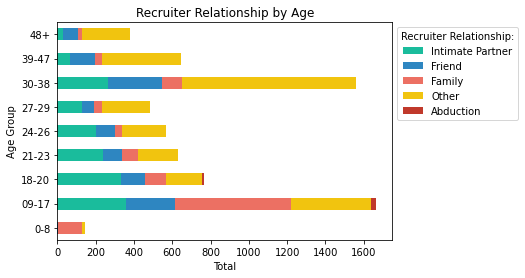
Once our data was cleaned, country name and regional data was merged with our data set on both citizenship country code and exploit country code. The country names and regional data was then used during our analysis.

1. **How does someone end up being trafficked?**

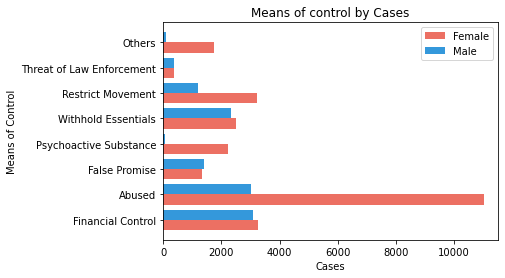
We wanted to know if there was any correlation between the Recruiter Relationship and the victims. For the majority of the cases reported, the recruiter did not have any initial relationship with the victims. The second most reported case was that the victims had an Intimate Relationships with their Recruiter. People who are between the age of 9-17 and 30-38 are most likely to be trafficked, this could be due to the type of exploitation related to this age group.



* Intimate Partner: The main takeaway is that teens and young adults(ages: 9-20) had higher numbers of cases with the recruiter being their intimate partner.
* Friends: highest cases are for Age group 9-17 with 256 cases and age group 39-47 with 279 cases with recruiter relationship being their friend.
* Family: highest cases are seen for ages group 9-17 with 605 cases that Family as the form of recruitment. The Cases slowly drop until age 30-38 where we see a spike of 107 cases.
* Abduction: Out of 39 cases reported in the whole data, 26 of the cases belong to the age group 9-17.
* Other: **Others DOES NOT MEAN UNKNOWN OR NO DATA.** It means that the type of relationship doesn’t fall under any of the specified categories or there was no type of relationship between the recruiter and the victim. Highest recruiter relationship with 2886 cases regardless of age.

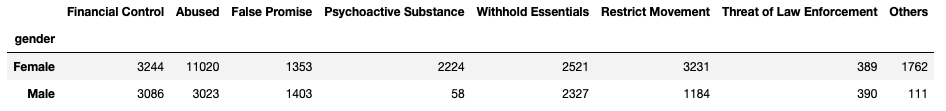


We also wanted to know how the traffickers were controlling the victims. For better understanding we grouped the means of control by the following categories: Abused (Physical, Psychological, Sexual, Threats), Financial Control (Debt Bondage, Restrict Financial Access, Excessive Working hours, Earning Control), False Promise, Psychoactive Substance, Withhold Essentials (Withhold documents, Necessities, restrict medical care, control using children), Restrict Movement, Threat of Law Enforcement, and Others. Other Categories are means of control that was not specified in the data.



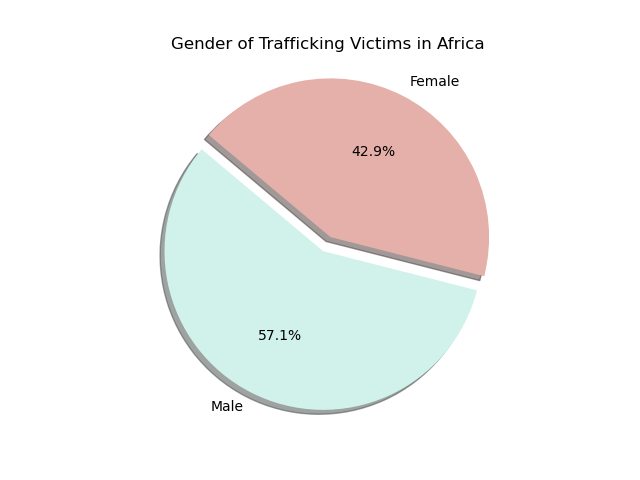
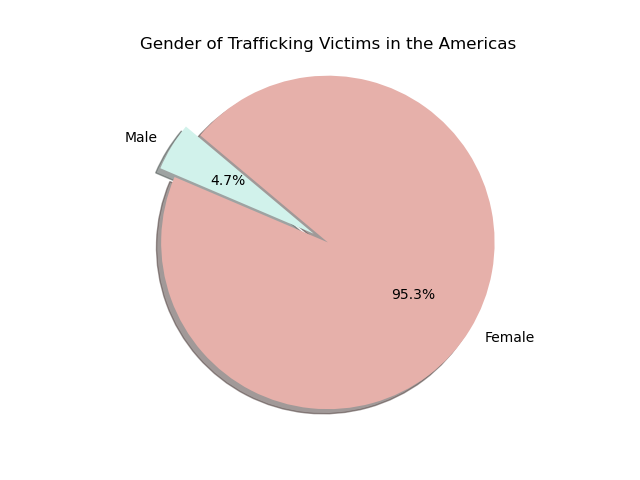
In our analysis, the highest means of control for females was through a form of abuse (11020 cases reported). For males, the highest means of control was through Financial Control (3086 cases reported). The lowest form of control reported for males was Psychoactive Substance with only 58 cases.

Side note: this also tells us that more females are trafficked according to the cases reported.

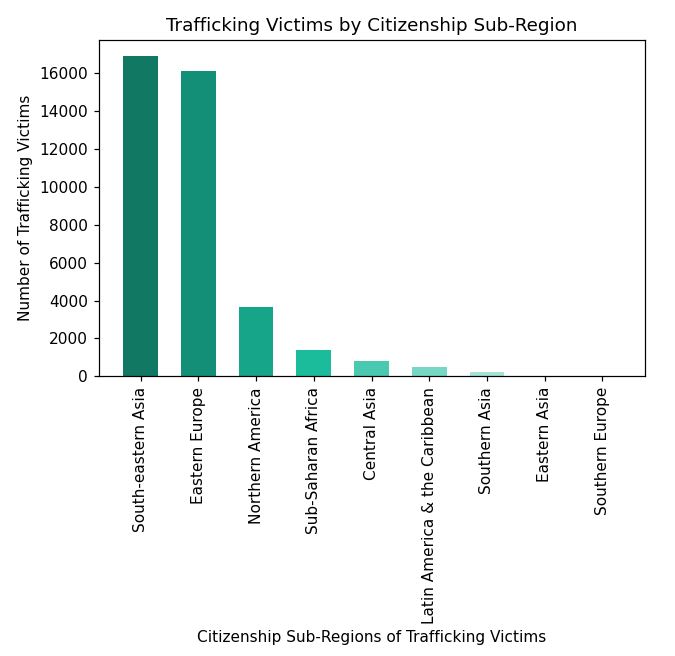
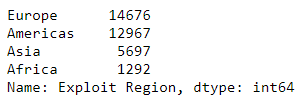
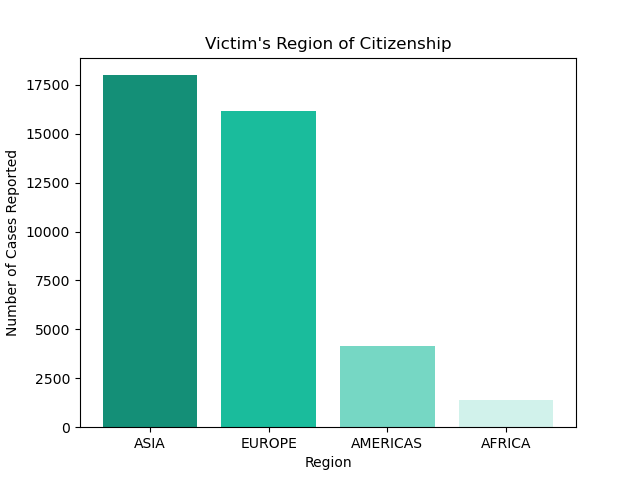
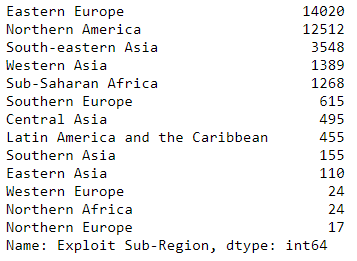


1. **Where are they from? Where are they being exploited?**

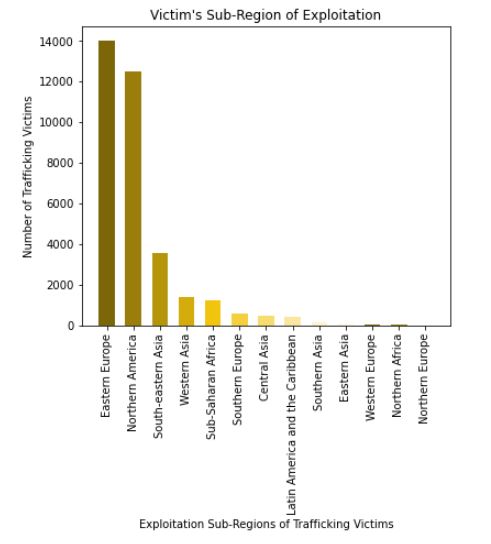
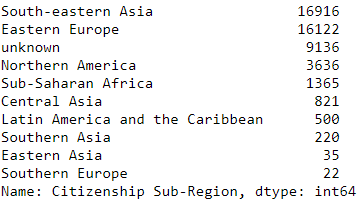
The majority of trafficking victims are citizens of Asia and Europe and they are being brought to Europe and the Americas for exploitation. One thing the data showed us is that the majority of trafficking victims were female in each of the regions *except* Africa where the majority of trafficking victims were male at 57.1%. Something to note is that the majority of trafficking victims in the Americas were female at a whopping 95.3% of total victims trafficked.

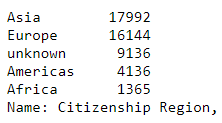
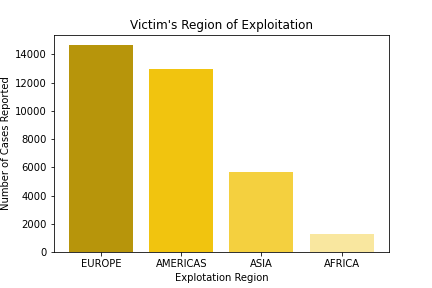
 

**Case counts by citizenship region and sub-region:**



**Case counts by exploit region and sub-region:**

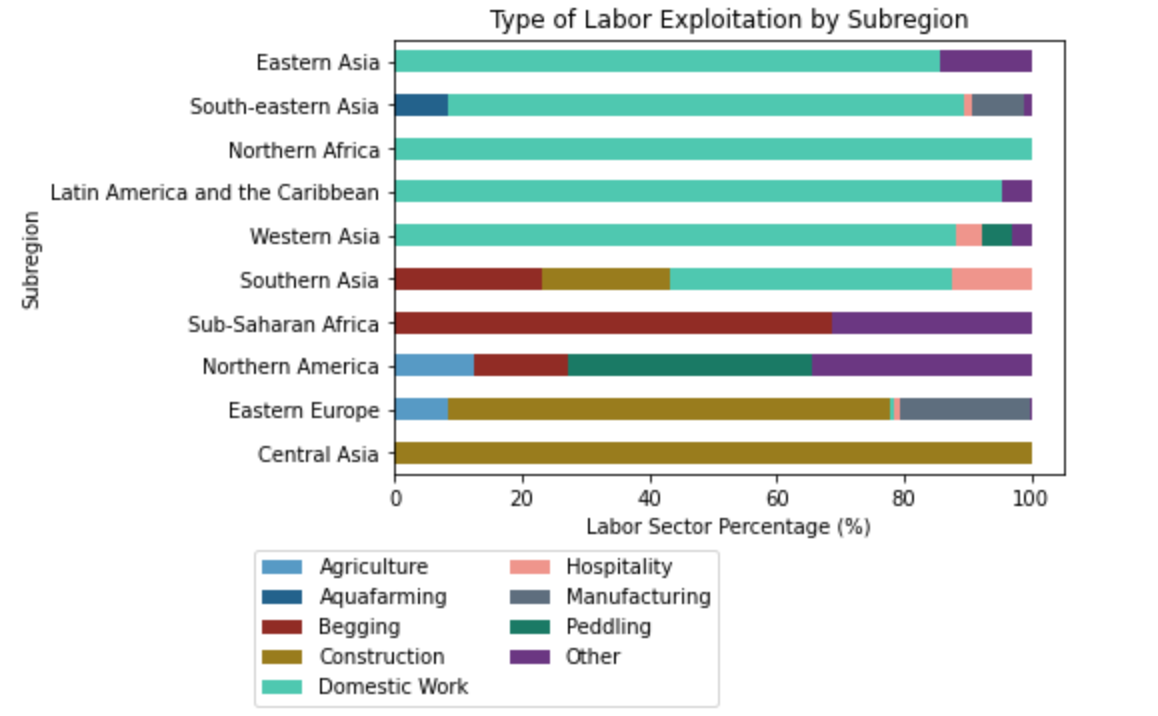




1. **In which industries are trafficked persons most likely to be exploited?**

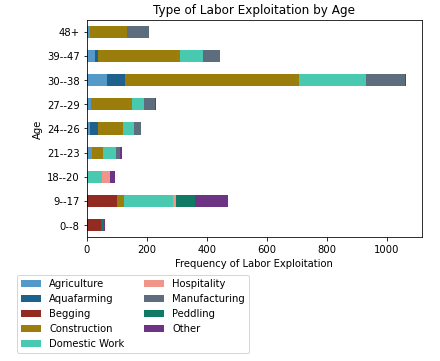
We wanted to understand the industries behind the human trafficking crisis, and the demographic characteristics for each industry involved. Overall, domestic work represents the majority of categorized exploitation cases, followed by construction and manufacturing.

Figure 3.1:



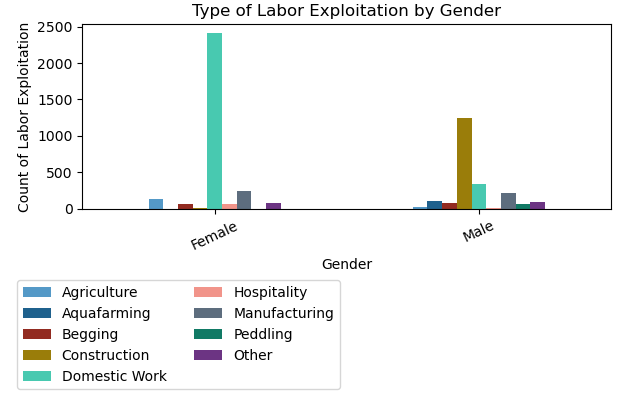
Central Asian and Eastern European labor exploitation is dominated by the construction industry. Domestic work is the most prominent industry category in Southern, Western, and Eastern Asia, as well as Northern Africa, and Latin America and the Caribbean. Sub-Saharan Africa has the highest proportion of begging, whereas Northern America has the most diversity within the sectors of labor exploitations, including agriculture, peddling, and a high rate of “other” reported.

Figure 3.2:



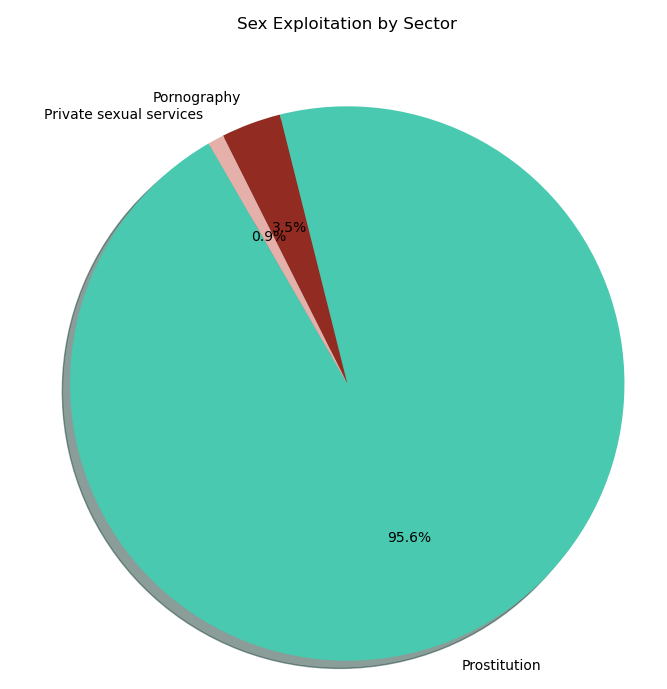
As for age groups, we see that traditional industries like agriculture, manufacturing, and especially construction skew toward the older groups, with a peak in overall exploitation for victims in their 30s. For the two under-age subgroups, we see a high rate of non-traditional labor sectors such as begging and peddling.

Figure 3.3:



There is also a gender bias when it comes to the dominant industries of trafficking. Female trafficking victims overwhelmingly are exploited as domestic workers, whereas male victims overwhelmingly end up in the construction industry. The gender distribution is roughly similar in the categories of begging and manufacturing.

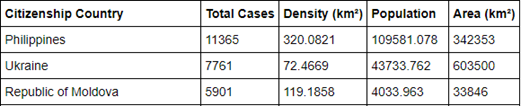
Figure 3.4:

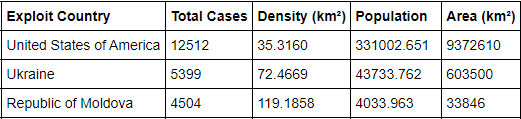


When looking at the category of sex exploitation, the prostitution industry is overwhelmingly dominant, capturing 95.6%, with pornography industry coming second at 3.5%, and private sexual services third at 0.9%. This dominant proportion in prostitution is consistently strong across subregions and age/gender demographics.

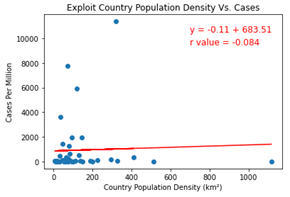
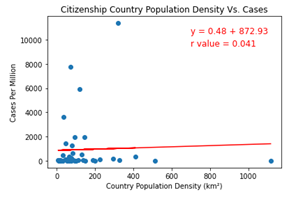
1. **Describe data in terms of population size**

Initially, we were interested to see if there is a correlation between country population size and number of human trafficking cases; however, we determined this isn’t a fair unit of measure due to the need to control for the disparity in country land area. In order to control for this, we pulled in data on country population density (population divided by land area) and merged it with our base clean data set to show the below tables:

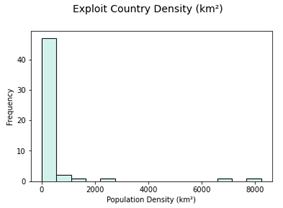
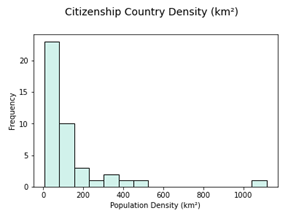




To see if there was a correlation between country population density and number of human trafficking cases, we used scatter plots to display the Citizenship & Exploit Country Population Densities and corresponding number of cases. Due to massive outliers, the r values were skewed and indicate little correlation between the two variables.



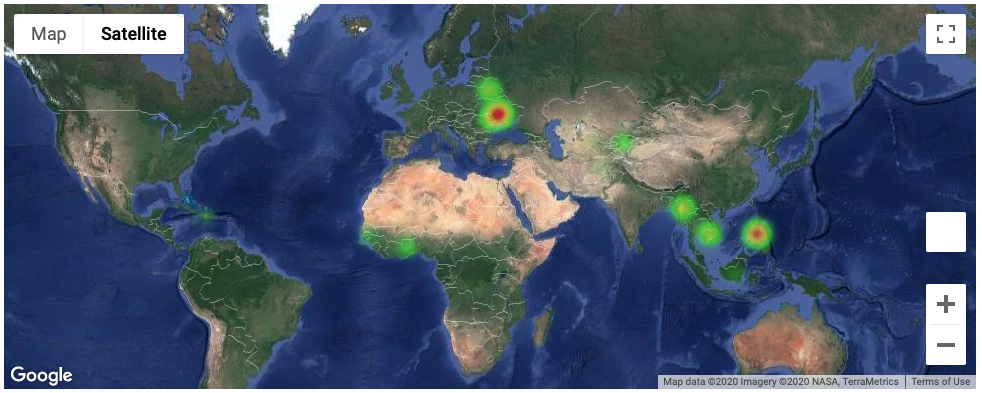
We then used histograms to see the distribution and frequency of country population density for Citizenship and Exploit Countries. We found that there are a great deal more countries that have comparatively smaller population densities per kilometers squared, meaning that there could be a correlation between the likelihood of being human trafficked and living in a less populous dense country; however, it should be noted that we were limited by our dataset size, as there were only 45 Citizenship Countries & 58 Exploit Countries.



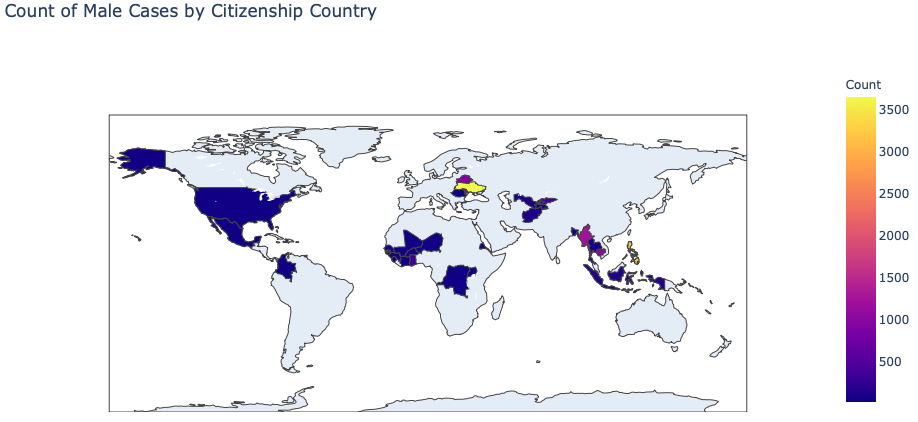


1. **Demographic Regional Comparisons**

During our analysis, we wanted to explore the number of cases based on citizenship country and exploitation country. And we wanted to group the data by gender. While charts are helpful to convey this information visually, we also wanted to incorporate some API calls and google maps we had learned. Using an API call we determined the latitude and longitude of country, region, and sub-region. Using the latitude and longitude, we started to explore heat maps based on count of cases in regions. The regional heat map did not add much value so we changed to citizenship country and exploit country.

  
Map 1: Heat Map of Male Cases by Citizenship Country

While the heat maps were effective as showing the ‘hotspots’ we felt that some of the data was lost. We looked into other map options and decided to use choropleth maps. We felt this better represented the data on a map without losing meaning.



Map 2: Choropleth Map of Male Cases by Citizenship Country

Lastly, we started to explore if citizenship country impacted exploit country. We began this analysis by determining the citizenship country with the highest number of cases where the exploit country was known. Based on our dataset, Ukraine was the country with the highest number of cases with known exploit countries. Then we displayed the exploit country using a marker layer on a google map over a heatmap. By doing so we see that the majority of those trafficked are kept within the same country.

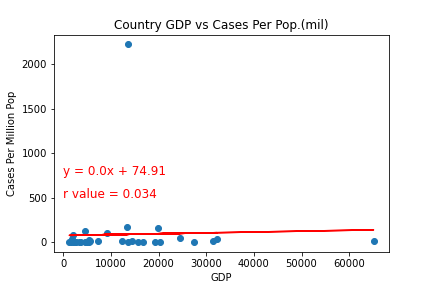
  
Map 3: Display of Where Ukranian Citizens are Exploited

## **6. Economic factors - correlation between poverty & trafficking**

Our final area of exploration was determining the relationship between a country’s economic factors and the probability of becoming a person who is trafficked. First we gathered information on the gini index which measures income inequality in a country. Unfortunately, most of the countries did not have any entries for the gini index. Therefore, we abandoned that comparison.

Next, we downloaded the per capita gross domestic product in terms of purchasing power parity to see if there was any correlation between a citizen's probability of being trafficked versus the wealth of the nation. We performed a linear regression analysis and plotted the results in the scatter plot shown below

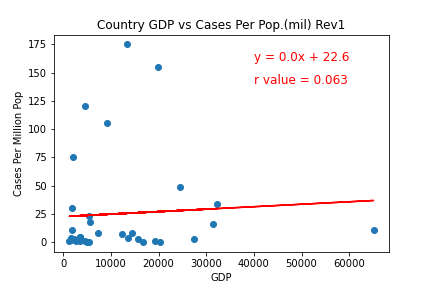
Figure 6.1:



Sources: CTDC/World Bank

The first mapping showed an obvious outlier. The Republic of Moldova showed very large cases per million population. If we had more time, we would investigate in-depth what is happening in the Republic of Moldova. Due to our time constraints we Reran the linear regression with the Outlier removed.

Figure 6.2:

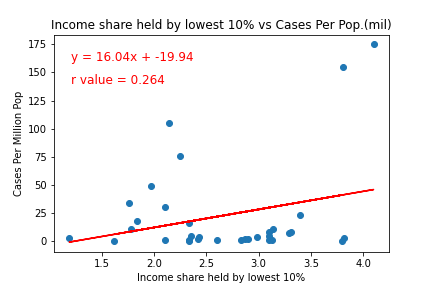


Sources: CTDC/World Bank

The results of the linear regression with the outlier removed showed there was little to no correlation between the probability of Being Human trafficked and the GDP of a country. With an r value of 0.063, we concluded the data showed no measurable correlation between the datasets

Since the wealth of the nation, measured as GDP, showed no correlation to the probability of being trafficked; We turned our attention to measures of poverty within a country. Linear regression analysis was performed on our next data set; income share held by the lowest 10%. he hypothesized income inequality would relate to cases per million, the results showed weak or no correlation with an r value = 0.264.

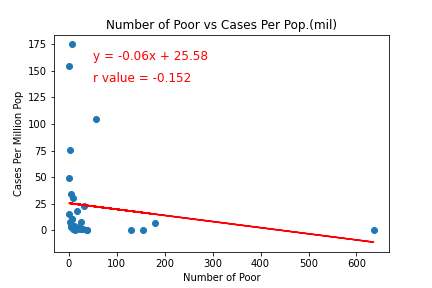
Figure 6.3:



Sources: CTDC/World Bank

Undaunted, We performed one more linear regression. This time the number of poor in a country was compared to the cases per million population. Again, linear regression produced an extremely weak correlation. This time there was a very weak negative correlation at -0.152.

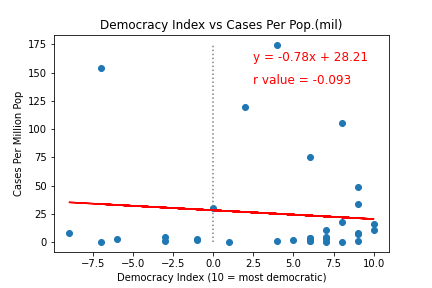
Figure 6.4:



Sources: CTDC/World Bank

Finally, we examined the relationship between the level of democracy in the country and the number of cases per million. The Democracy index measures perfect democracy at 10 and a total autocracy at -10. The results of linear regression for the two data points revealed a very weak to no correlation between the data points.

Figure 6.5:



Sources: CTDC/World Bank

**Summary**

We found that the primary method of human trafficking was through a recruiter with no prior relationship with the victim, and the age ranges of 9-17 and 30-38 have the highest incidences of human trafficking risk. The primary means of control to keep people within the system of exploitation is through some form of abuse for female victims and through financial control for male victims.

When considering the flow of humans in the global network of human trafficking, the majority of victims are citizens of Asian and European countries who are taken to European and American countries for exploitation. In the case of Ukrainian trafficking victims (the citizenship with the highest known cases), we found that almost all destination countries were nearby Eastern European and Central Asian countries.

With regards to the industrial data within each subregion of the global network, we found that the domestic work sector of labor exploitation dominated across South Asia, East Asia, the Middle East, and Latin America. In Central Asia and Eastern Europe, the construction industry is responsible for most cases. When considering age, the older age groups, especially 30-39, were most likely to be exploited by traditional sectors like construction and manufacturing, whereas under-age groups were more exploited by non-traditional sectors like begging and peddling. In terms of gender, the vast majority of female victims were exploited via domestic work, whereas the vast majority of male victims were exploited by the construction industry.

We tested multiple correlations of human trafficking to economic and democratic indicators for our target countries and found no correlation. This implies that the economic strength or democratic characteristics of a country offer no protections against the risk of human trafficking. However, we must consider the primary limitation with our dataset, which is that each country will have its own trafficking reporting bias, with no standardization across the world.