Methods of Preprocessing Sparse Human Trafficking Data

Adam Dameron

April 27, 2023

Abstract

This research is on the methods of preprocessing a large human trafficking data set in a manner that is conducive to predictive modeling. These methods include ways to account for sparseness the data, and ways to augment additional data such as the development of a country, so that models can be created for the purposes of enforcing and creating policies to prevent human trafficking on a global scale.

Questions For Dr. Duncan

- Should I use ”we”, ”I”, or non-firstperson words when describing the methods used?
- Should I move the really long table to an appendix?
- Should I put all tables and graphics in an appendix for quick reference?

1 Purpose/Background

Human trafficking occurs on a global scale, and has been recorded as an international concern since the year 1913 (Aromaa 2007). Many attempts have been made at defining it, yet these definitions remain unclear and vary greatly. Human trafficking is strongly believed to occur in every country, and victimizes individuals of all ages, genders, and backgrounds (Jac-Kucharski 2012).

A major gap in the global understanding of international human trafficking, is in how human trafficking is affected by changes in a country’s economic and social development. The United Nations mentions this fact in its ”Introduction to Human Trafficking,” and describes trafficking as a ”multidimensional problem,” where a development is applicable (Kangaspunta 2008). However, there exists very little quantitative analysis into how trafficking and development are connected. Additionally, it has been shown previously that economic and social development has a substantial effect on the migration rates into and out of a country, which in-turn is correlated with human trafficking (Litan 2000; Omar Mahmoud and Trebesch 2010). This research seeks to provide more insight into what connections exist between development and human trafficking, and quantify these relationships in a manner which is easy to interpret.

1.1 Background Information

The trafficking of humans is a multifaceted problem which exists at all levels from local to international (Aromaa 2007; Jac-Kucharski 2012). The United Nations defines human trafficking as “recruitment, transportation, transfer, harboring or receipt of people through force, fraud or deception, with the aim of exploiting them for profit” (Raymond 2002). In order to truly understand the scope of this problem, one must begin by gaining an understanding of the different types of trafficking, and the reasons why it is such a widespread issue in all locations; regardless of economic, social, and cultural stability.

Defining Types of Trafficking

The United States State Department has has a designated ”Office To Monitor and Combat Trafficking in Persons” since October of 2001. This office publishes a yearly ”Trafficking in Persons Report,” and in this report, two types of human trafficking are reported on: sex trafficking, and labor trafficking (*Understanding Human Trafficking* 2022).

The U.S. State department defines sex trafficking as ”...activities involved when a trafficker uses force, fraud, or coercion to compel another person to engage in a commercial sex act...” Similarly, labor trafficking is defined as ”...activities involved when a trafficker uses force, fraud, or coercion to exploit the labor or services of another person (*Understanding Human Trafficking* 2022).” Needless to say, these definitions lack depth, and seem vague in nature.

Polaris is a non-profit organization that reports, analyzes, and educates others on the intricacies of human trafficking in the United States. Alternatively to the U.S. State Deparment’s two proposed types of human trafficking, Polaris has defined twenty-five different types of human trafficking that are very-much present in the United States. To preface the 80 page report on ”The Typology of Modern Slavery,” Polaris states: ”...the ways humans are exploited differ greatly. Each type has unique strategies for recruiting and controlling victims, and concealing the crime. (Polaris 2017)”

The types of trafficking described by Polaris were based off of over 42,000 reports made to the ”Human Trafficking Hotline.” The types mentioned range from agricultural work and farming to personal sexual servitude and escort services. However, Polaris notes that most cases of human trafficking will involve more than one type, and even these types may not fit neatly into ”sex trafficking” or ”labor trafficking” (Polaris 2017).

In conclusion, there are no neatly defined categories for human trafficking, except the general consensus in academia, law enforcement, and non-profits, seems to be that aside from a few edge cases, most instances of human trafficking unfortunately have to be reduced to being in one of two major categories, as any deviation from this results in inconsistent and contradictory definitions. It is for these reasons that a different method of classifying human trafficking cases will be attempted, but is important to recognize the work already done on this topic by powerful entities such as the U.S. Department of State, and a large organization such as Polaris.

United States's Involvement in International Trafficking

The United States has long been growing into the global economic leader it is today (Litan 2000). With this increase in global influence comes an increase of international crime within the country's borders. Sex and labor trafficking have both been long-standing issue within the United States. In 2000, the United Nations developed the Palermo Protocols; These protocols sought to ”prevent, suppress and punish trafficking in human beings” (United Nations General Assembly 2000). In 2005, the United States ratified the protocols in an effort to fight the increasing prevelance of labor and sex trafficking.

Despite both labor and sex trafficking being present in the United States, most media and law enforcement attention has specifically been targeted towards sex trafficking, particularly of children (Wallinger 2010). Logically it can be assumed that given the influence the US has on global trade and international crime, that there is are instances of individuals of other citizenships being exploited within the country, and in ways that are not just sex trafficking.

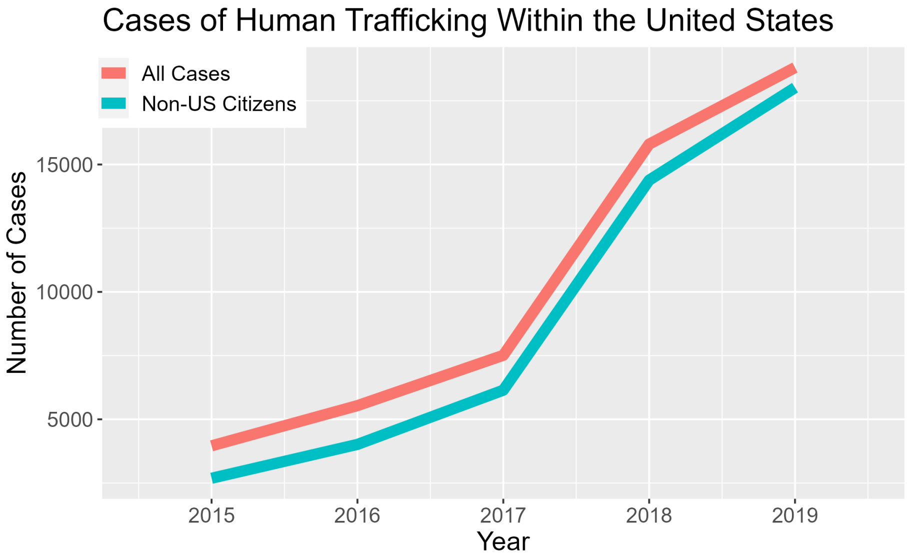


Figure 1: Created with data from *The Global K-Anonymized Dataset* 2021

As evidenced in this simple chart, in a span of only 4 years, the number of cases in the United States alone has nearly quadrupled. Although, it is important to note that this is likely to be a severe under-counting of the true number of human trafficking victims each year. Even so, having roughly 20,000 cases in 2019 alone is rather extreme, and shows that this is very much an issue that exists in the United States. Of additional importance is the fact that a large proportion of all cases in the US are ones involving non-US citizens. This means that not only is trafficking present in the US, but international trafficking is of major concern.

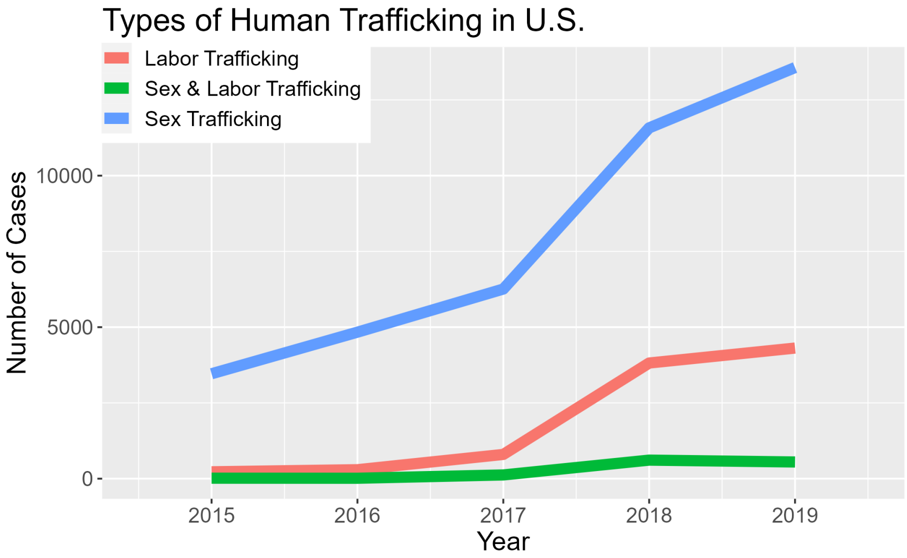


Figure 2: Created with data from *The Global K-Anonymized Dataset* 2021

From this plot, one can see that again, each year there is an increase in the total number of reported cases; however, the plot also shows that a significant number of cases exist in which sex trafficking is not the sole type of trafficking present. Thus, even within th U.S., international trafficking is a very real issue, and it exists in all forms, not just sexual. Thus it would be beneficial to better understand, or even predict, what aspects of these victims law enforcement should be knowledgeable of.

Scale of International Trafficking

It is difficult to put into words the scale of international human trafficking. However, by looking at the number of human trafficking victims reported in each country, and then determining what percent of them are from other countries, the scale of the problem can be visualized.

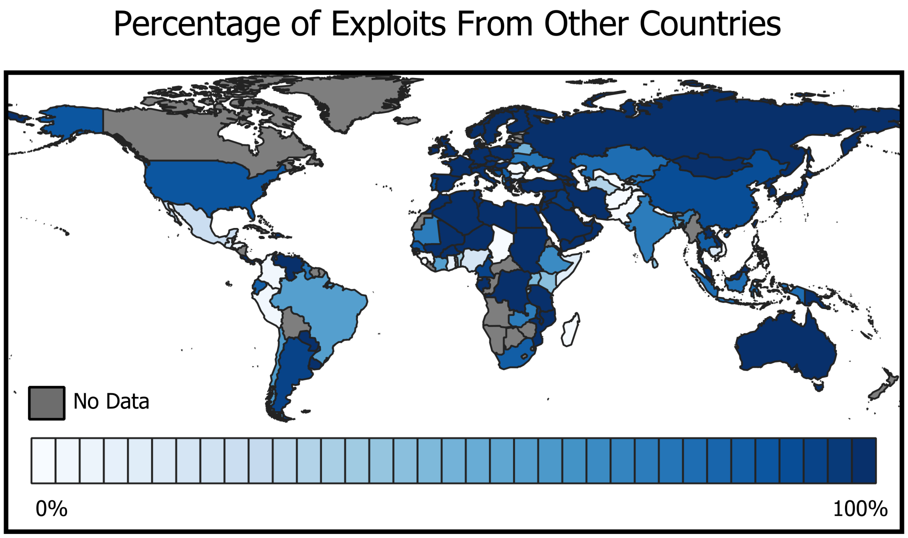


Figure 3: Created with data from *The Global K-Anonymized Dataset* 2021

In this map, each country is colored one of many shades of blue, where darker shades correspond to a higher percentage of international victims within the country. One can see in the map that there are a lot of dark-blue regions. Some distinctly blue regions are North-Africa and Europe. There are also some scattered regions in South-East Asia which have a high percentage of international trafficking victims.

Of additional interest are the very light-blue or even white regions of the map. these regions have a low percentage of international trafficking victims, which means that these countries almost exclusively have victims who originate from within the country. While this visualization gives us an idea of the prevalence of international victims who are exploited in each country, it does not show us how many victims originating from a country are ”exported” to other regions.

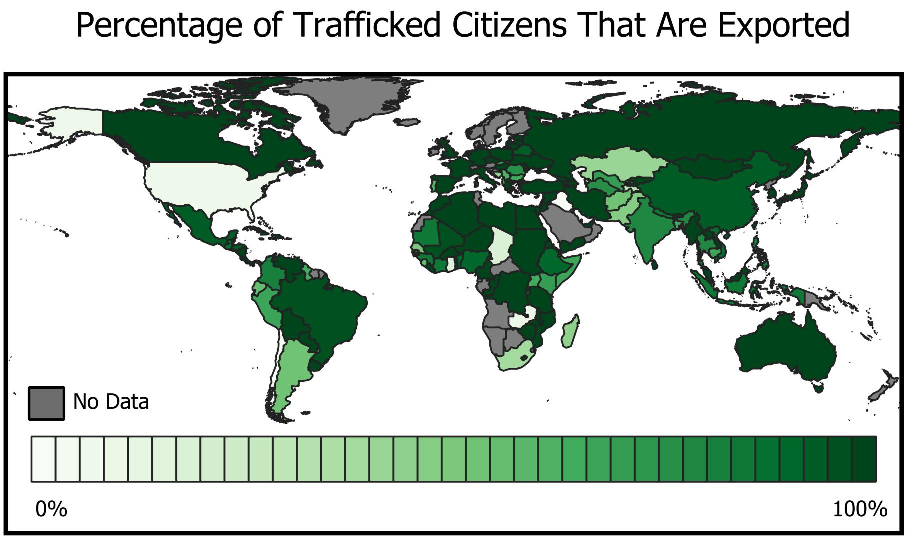


Figure 4: Created with data from *The Global K-Anonymized Dataset* 2021

This map is similar to the first one in that a darker color (in this case green), indicates a higher percentage of trafficked citizens who are exported from their country of origin. These two maps on their own do not paint a good picture, but when viewed together, they provide useful qualitative information.

Some preliminary conclusions one could make from these maps. For example: Mexico has a low percentage of victims from other countries, and a large percentage of Mexican citizens who are exploited, exploited in other countries. Thus it can be inferred that Mexico is an "exporter" of human trafficking victims.

The Northern region of Africa has a high percentage of citizens that are exported, as well as a large percentage of exploits which are from other countries. This means that countries in the region displace many of their exploited citizens to other countries, but bring in victims from other countries. This suggests that individuals in North Africa are trafficked to other regions to fill a certain "role," and victims from other regions ar imported to fill a different role. In fact, this is shown to be true as a result of previous studies on the region.

The smuggling of migrants and laborers from Northern Africa to Europe explains the high percenage of exploited citizens that are exported (Malakooti 2016). The high percentage of victims within these countries coming from other regions is explained by the trafficking of individuals from West Africa to North Africa. Many West African migrants attempting o get to the Mediterranean Sea, are often intercepted and trafficked along their route through Algeria and Libya (Sanchez et al. 2018).

Mexico and North-Africa are just two examples of the many types of qualitative conclusions one can make from these maps; however, a major question these maps could never answer is what types of trafficking are present, what indicators exist to determine the typology of trafficking victims in a region, and most importantly, how the stage in development of a country effects these indicators and types of trafficking.

1.2 Objective

The Global K-Anonymized Dataset (CTDC Dataset) is a dataset which was developed and is maintained by the International Organization for Migration (IOM). This dataset was created with the aim of providing a large centralized database of human trafficking cases all over the world. This dataset contains over 150 thousand entries, and has data from 189 countries. This collection of data exists publicly, and each entry has been anonymized in order to protect the identities of the victims.

The CTDC dataset was initially created to solve the problem of human trafficking databases being inconsistent, and unreliable. By creating this dataset, IOM created a database which is easy to use and navigate. The creation of this dataset has solved a major issue of quality data analysis on human trafficking being impractical, or even impossible.

The objective of this research is to explore ways of preprocessing the CTDC dataset in a way that fosters the future development of models that show what types and indicators of trafficking are present in countries of different stages of development. This serves to fill is a significant gap in the existing knowledge offered to law enforcement and law makers in regards to international human trafficking. By creating these models, it will provide these entities with information more specific to the country they serve, and will allow them to make more informed decisions.

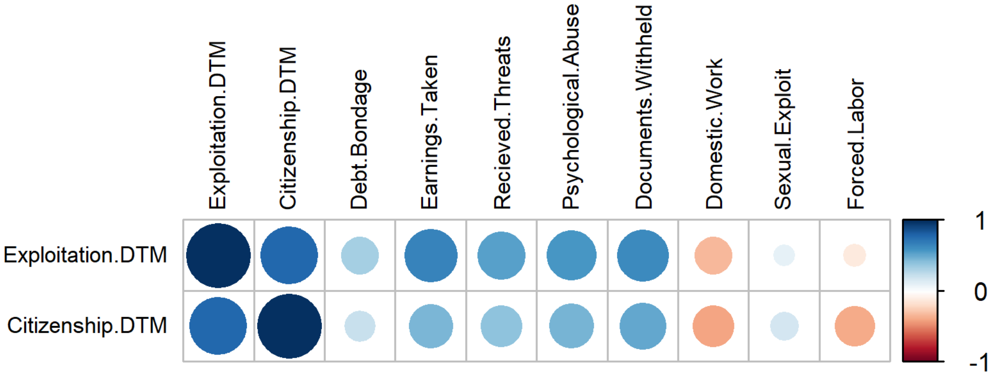
2 Preliminary Reasoning and Analysis

The Counter Trafficking Data Collaborative (CTDC) has a publicly available, anonymized dataset, which is a collection of over 400,000 entries for human trafficking victims (*The Global K-Anonymized Dataset* 2021). The database has plenty of useful information, such as country of citizenship, country of exploitation, and many variables expressing the type(s) of trafficking the victim experienced. This data set is the focus of this research, as it is the largest publicly available database for instances of international human trafficking.

While the CTDC database is extremely large, it does not contain information on a country's development. For this, one needs a numerical way to categorize countries into "stages," as the goal is to make a generalized model for stages of development, not one that pertains to specific countries. One way to categorize countries would be by using the Demographic Transition Model (Bongaarts 2009). Using this model, a country can be categorized into one of 5 stages, with stage 1 being the lowest developed countries, and stage 5 being the most developed.

Unlike many other ways of classifying countries in terms of their development, this model uses birth, death, and migration rates to determine a country's stage (Bongaarts 2009). While these rates are not directly related to a country's economic growth or decline, the model has been shown to be very successful at expressing economic changes, while not being as strongly influenced by sudden, temporary economic setbacks or "booms" (Kirk 1996; Bar and Leukhina 2010; Galor and Weil 2000). The model has even been shown to help predict changes in the "shadow economy," in areas of illegal activity such as drug trafficking (Schneider 1994). It is for these reasons that I believe there exists a way to efficiently model and predict changes in the human trafficking typology of a country as it becomes more or less developed.

For our data, a complete entry is a case in which a value is recorded for all variables in the data set. An incomplete entry is any entry in which there are values which are unknown or missing, such as the age of the victim, or the type of trafficking experienced. By looking at complete entries in the CTDC Dataset, a preliminary correlation comparison can be completed. One can take both the stage of the citizenship country, and exploitation country for a victim, and see if this correlates with any aspects of trafficking they experienced. The closer to perfectly correlated (-1 or 1) or connected two variables are, the more red or blue the circle will be. If a circle is red, that means the trait is seen more in low stage countries, and if it is blue, it is observed more often in high-stage countries.



Each of the variables included in the plot have at least a moderate correlation with DTM. As such, it would be beneficial to further quantify and model what types of relationships do exist within the data.

2.1 Why Demographic Transition Is Important

In *Human Trafficking: A Perfect Storm of Contributing Factors* (Tiano 2012) a connection is theorized between human trafficking and the Demographic Transition Model. The author, Susan Tiano, talks about how the rise and fall in birth and death rates has an effect on human trafficking. By first talking about the growth of the ”underground economy” in relation to shift as in Demographic Transition, Tiano sets the stage to infer that trafficking is no different. Tiano's research further validates the claim that Demographic Transition does have an effect on human trafficking, but Tiano was unable to quantify such a relation. Thus, being able to quantify these relationships would fill a major gap in the current knowledge of the human trafficking landscape.

In *Potential Impact of Climate Change on Human Trafficking* (Sheu et al. 2021), there is discussion of many of the direct and indirect effects that climate change has on international human trafficking. Of high interest in this article is what is said about human trafficking in the United States, in the aftermath of Hurricane Katrina. According to this article, the Federal Government temporarily removed certain worker protections in the New Orleans area after the hurricane. This led to a huge influx of labor trafficking victims into the region in order to help rebuild. This plainly demonstrates that labor trafficking can be used to fill a sudden demand for cheap labor (Sheu et al. 2021). It also is an example of the major, somewhat instantaneous effect that federal governmental legislation can have on the typology of trafficking within their borders. Thus, making informed lawmaking decisions is vitally important in preventing human trafficking.

In the case of Hurricane Katrina, the Gulf Region essentially became less developed than the rest of the United States, and as a result, labor trafficking spiked (Sheu et al. 2021). This is yet another example of the effect that development can have on trafficking in a region.

3 Variables

The following table defines all of the variables in the data set. On the right is the name of the variable as it occurs in the data set and in this paper, and on the right is an explanation of what the variable represents.

Table 1: Variables related to human trafficking

Variable	**Definition**
yearOfRegistration	The year in which the victim of human trafficking was registered.
gender	The gender identity of the victim of human trafficking.
ageBroad	The age range of the victim of human trafficking.
majorityStatusAtExploit	Whether the victim of human trafficking was a minor or an adult at the time of exploitation.
citizenship	The country of citizenship of the victim of human trafficking.

CountryOfExploitation	The country where the victim of human trafficking was exploited.
traffickMonths	The number of months the victim of human trafficking was trafficked.
meansOfControlDebtBondage	Whether the victim of human trafficking was subjected to debt bondage as a means of control.
meansOfControlTakesEarnings	Whether the victim of human trafficking's earnings were taken as a means of control.
meansOfControlThreats	Whether the victim of human trafficking was subjected to threats as a means of control.
meansOfControlPsychologicalAbuse	Whether the victim of human trafficking was subjected to psychological abuse as a means of control.
meansOfControlPhysicalAbuse	Whether the victim of human trafficking was subjected to physical abuse as a means of control.
meansOfControlSexualAbuse	Whether the victim of human trafficking was subjected to sexual abuse as a means of control.
meansOfControlFalsePromises	Whether the victim of human trafficking was subjected to false promises as a means of control.
meansOfControlPsychoactiveSubstances	Whether the victim of human trafficking was given psychoactive substances as a means of control.
meansOfControlRestrictsMovement	Whether the victim of human trafficking was subjected to restrictions on movement as a means of control.
meansOfControlRestrictsMedicalCare	Whether the victim of human trafficking was denied access to medical care as a means of control.
meansOfControlExcessiveWorkingHours	Whether the victim of human trafficking was forced to work excessively long hours as a means of control.
meansOfControlThreatOfLawEnforce	Whether the victim of human trafficking was threatened with law enforcement action as a means of control.
meansOfControlWithholdsNecessities	Whether the victim of human trafficking is deprived of basic needs such as food, water, and medical care in order to coerce them to comply with their trafficker's demands.
meansOfControlWithholdsDocuments	Whether the victim of human trafficking has their identification documents or other legal papers confiscated by their trafficker, making it difficult or impossible for them to leave or seek help.
meansOfControlOther	Whether any other means of control that a trafficker may use to exert power and control over their victim, such as blackmail, isolation, or physical restraints.
isForcedLabour	Whether the victim is forced to work against their will under the threat of punishment or harm.

isSexualExploit	Whether the victim is forced to engage in sexual activities against their will. This can include prostitution, pornography, or other forms of sexual exploitation.
isOtherExploit	Whether the victim is forced to work or perform services against their will in a situation that does not involve sexual exploitation or forced labor. This can include domestic servitude or begging.
typeOfLabourAgriculture	Whether the type of labor the victim of human trafficking was subjected to was related to agriculture.
typeOfLabourConstruction	Whether the type of labor the victim of human trafficking was subjected to was related to construction.
typeOfLabourDomesticWork	Whether the type of labor the victim of human trafficking was subjected to was related to domestic work.
typeOfLabourHospitality	Whether the type of labor the victim of human trafficking was subjected to was related to the hospitality industry.
typeOfLabourOther	Whether the type of labor the victim of human trafficking was subjected to was was not covered by the other categories.
typeOfSexProstitution	Whether the type of sexual exploitation the victim of human trafficking was subjected to was related to prostitution.
typeOfSexPornography	Whether the type of sexual exploitation the victim of human trafficking was subjected to was related to pornography.
typeOfSexOther	Whether the type of sexual exploitation the victim of human trafficking was subjected to was was not covered by the other categories.
recruiterRelationIntimatePartner	Whether the recruiter of the victim of human trafficking was an intimate partner or spouse.
recruiterRelationFriend	Whether the recruiter of the victim of human trafficking was a friend or acquaintance.
recruiterRelationFamily	Whether the recruiter of the victim of human trafficking was a family member.
recruiterRelationOther	Whether the recruiter of the victim of human trafficking had a relationship with the victim that was not covered by the other categories.

4 Results

This section serves as a summary of the methods used, as well as the results which were obtained.

4.1 Removing Variables

Section 5.1 explores the possibility of addressing the issue of missing data by analyzing where the missing data occurs, and removing variables which appear to be insignificant. By removing these variables, the number of complete rows, or records in which there was no missing data, was increased. The big hurdle with this method was determining which variables should be removed. There were plenty of variables, mostly in the means of control section of variables, which had a very large number of missing values. These variables contributed significantly to the overall number of missing values (see figures 5 and 6) removing them would increase the number of complete entries in the data set, but it is difficult to ensure important information is not lost. As such, there were attempts to find a balance between removing insignificant variables, and increasing the number of complete entries in the data set. This was done with the use of correlation analysis in section 5.1.

Section 5.1.1, starts by removing variables that are known to not be beneficial in the modeling phase based off of the variable definition alone. These variables include text variables that summarized other data and variables would not be significant in a predictive modeling setting. "Data Source" was one such example of this, as it provided information on how the case was reported and would not necessarily provide any information about the case of trafficking itself.

Next, in section 5.1.2, quantitative methods to remove variables were explored. This was achieved by looking at things such as frequency of missing values within each variable, and combinations of variables which were frequently missing together.

4.2 Means Of Control Frequently Left Unreported

In section 5.1.2, the variables had the highest frequency of missing values were analyzed. By ordering the variables by the frequency of missing values, there were no variables that were missing between 70,000 and 80,000 (see figure 5). From this, variables missing at least 80,000 entries (82% of total entries) were analyzed. Each one of these variables was in the "meansOfControl" variable category. This showed us that there was a pattern of "meansOfControl" variables having large amounts of missing data. As a reminder, the meansOfControl variables are variables which give binary indicators for wheather or not a certain type of control was used on the victim.

Analyzing the meansOfControl variables, showed that over 60,000 entries in the data set had no meansOfControl variable that was observed. Meaning that these entries had only values of 0 and missing for all meansOfControl variables. This was interesting, as it indicated to us that these variables are frequently left unrecorded. The way in which an individual or group is being controlled throughout the course of them being trafficked is an important factor to consider when finding ways to save the victims, or stop the crime entirely. It was determined that in the process of creating a complete data set, that these variables would not provide much useful information. The sheer lack of information contained within these variables, as is evidenced in 7, led to the decision to remove these variables from our data set.

4.3 Inconsistent Data Entry Practices Between Categories

After removing the meansOfControl variables, there still existed a significant problem of missing values. In figure 8 it is shown exactly where missing values exist within observations. While the missing percentages for each variable were initially analyzed, it became more evident that the missing values followed certain patterns. Within certain variable categories, there appeared to be significant correlation between one variable missing in a category and all variables also missing in the category. However, in some categories, one can see instances of some values missing and others being present. This led to the conclusion that data entry practices are inconsistent between countries and variables. For example, in recruiter relation, it was observed that the entire category is present or missing. This means that if the recruiter relation is known, then a 0 is recorded for all other recruiter relation variables. Additionally, if the recruiter relation is not known, then all values are missing. This is in contrast to the means of control variable, as was shown before.

The entries in our data set had no means of control recorded (no value of 1 for any variable in the category) roughly 66% of the time. This means that 34% of the time, there was at least one means of control that was observed. However, it was shown that there are many variables in this category which are missing over 82% of the time. Since means of control is known 34% of the time, and some values are missing 82% of the time, then one can conclude that some values are left as missing, despite knowing at least one means of control, at least 16% of the time. This value is small, but in recruiter relation, these instances of having a missing value recorded even though the recruiter relationship was known, happened zero times. This implies that means of control variables and recruiter relation variables are recorded differently. It can be inferred that more inconsistencies such as this are also present for other categories of variables.

4.4 Sex Trafficking Frequently Missing Variables

In section 5.1.2 it was explored which variables (outside of "meansOfControl") were most frequently missing together. Figure 9 was created as a means to visualize these interactions. From this, it was discovered that "isAbduction" (wheather or not a victim was abducted or kidnapped) and all of the "typeOfSex" (wheather or not a given type of sexual exploit was observed) variables were the most frequently missing group. There were roughly 25,000 entries in which just these 5 variables missing together. The next largest group had the same set of variables, but also included ”ageBroad”. This group had a similar number of instances in which those were the only variables missing. This is significant, as it shows that the ”typeOfSex” variables are frequently left as missing together. This leads to the conclusion that there is likely some external factor that causes these variables to all be unknown. This is interesting, as it implies that the investigating sex trafficking oftentimes does not lead to knowing the type of trafficking present. This exposes a severe gap in knowledge and enforcement techniques in the realm of trafficking.

4.5 Data Suitability After Removing Problem Variables

After removing the variables we thought to be insignificant due to inconsistent data entry, observational difficulty, or irrelevancy, the dataset had 6,672 complete entries and 22 variables. This felt like a suitable data set for modeling. It was less sparse than the original data set, and contained information that was easier to interpret after modeling. To ensure suitability for modeling, a correlation plot (figure 10) was created in section 5.1.5. This plot indicated that there were no instances of a variable being significantly correlated with all other variables. This also included the DTM Stages for both exploitation and citizenship countries, a quantifier of how developed a country is. Meaning that a countries stage in DTM was not correlated with every variable in the data set. However, the typeOfLabour variables seemed to have the least significant correlation with other variables. This implies that typeOfLabour is somewhat random for victims of labour trafficking. For example, if you know a victim was used in the agriculture indistry, that does not give you any indication of whether or not they were also involved in the hospitality industry.

It is important to remember that this is from a pre-processed data set, so it is not truly representative of the data. However, it is a data set which is believed to be the best contender for modeling. The correlation analysis (in section 5.1.5) did not reveal any obvious problems with the data set, so this data set was used for further analysis.

4.6 Over-representation of Stage 5 Countries

In the final pre-processed data set, an issue was discovered of one country making up the majority of stage 5 citizenship countries (see section 5.1.6), and a separate country over-represented in stage 5 exploitation countries. Further analysis showed that Ukraine was over-represented as a citizenship country, and Russia was over-represented as an Exploitation country. This could be an issue, but it was also possible that there was enough variation between entries of these countries, that they would be indistinguishable from other stage 5 countries. However, a Chi-Square test revealed that both Russia as an exploitation country and Ukraine as the citizenship country, were significantly different from all other stage 5 countries in regards to every other variable. This by itself is an interesting result, and quantifies the significance of the human trafficking crisis facing Ukraine and Russia, as is outlined in *Human Trafficking: The Secret to*

Putin's Economy 2020. However, this result meant that any models created with the stage of demographic transition would be over-fitted to Russia and Ukraine. Thus this pre-processed data set was determined to not be useful.

4.7 Salvaging Missing Values

Section 5.2.1 began the process of including the missing values as a category for each variable. Each missing value in the data set was assigned a value of -1 to allow for a simple way to notate and keep track of which values are missing.

In the section, it is noted that there were many instances in which a value of 1 was recorded for a sub-variable within certain variable categories, and all other variables in the category were missing. The assumption was made that this was likely a mistake, and these missing values should actually be zeros (this is further explained in section 5.2.1). Therefore, each row in the dataset was checked, and instances such as these were "salvaged" by replacing the missing values with 0.

Then, table 4 presents a outlines the number of rows that were salvaged in each category of variables. The table shows that there were many instances in the data set in which one subcategory or type was observed, and others were left as missing. Then figure 11 was created to visualize the salvaged rows by country and variable category combination. The visualization shows that the most frequently salvaged rows are the ones involving Means of Control and Recruiter Relation. The United States is the largest contributing exploitation country to these instances of salvageable rows.

4.8 Minimization of False Positives

Type I and Type II errors are statistical concepts that can be applied to the context of human trafficking to evaluate the accuracy of identification and intervention efforts.

A Type I error, also known as a false positive, occurs when a person is incorrectly identified as a victim of human trafficking. This can happen if law enforcement, social services, or other stakeholders mistake other forms of exploitation for human trafficking or if they misinterpret benign situations as trafficking.

On the other hand, a Type II error, also known as a false negative, occurs when a person who is a victim of human trafficking is not identified as such. This can happen if the signs of trafficking are not recognized or if victims are reluctant or unable to come forward due to fear or mistrust of authorities.

When analyzing the rows that were salvaged in section 5.2.1, a conclusion was reached that the large number of salvaged rows in the United States could indicative of an intentional effort to minimize Type II Errors (false negatives). These errors occur when an aspect of trafficking is recorded as not observed or recorded but was actually present. It is potentially important in the context of investigating and combating human trafficking that these errors are minimized, as this prevents resources being allocated towards aspects of trafficking that are not present, before resources are used to combat the aspects of trafficking that are true positives (that were known to actually happen). It is difficult to determine that a certain aspect of trafficking did not happen, and making claims and decisions based solely off of an aspect of trafficking being recorded as 0 instead of NA does not make sense from a policy and enforcement perspective, as something being not-observed does not mean it did not happen.

As a result, missing values and values of 0 (not observed) could potentially be equivalent in a modeling setting. Since there is reason to believe at least one country recorded unobserved values as missing, it may be useful to apply this same reasoning to all rows in the data set. Ideally, the data would be representative of things which are known to be true, and not of things that are potentially true. Replacing all of the missing binary values with 0, brings the number of complete rows to over 22 thousand.

5 Problems of The CTDC Dataset

One of the most cited problems surrounding human trafficking data is the large amount of missing data entries (Tiano 2012; Polaris 2017), and this is especially true of the CTDC data set. The CTDC data set contains 63 variables and 0 of its entries are complete. In other words, each row in the data set has atleast one variable that is missing. This means that without any data manipulation, creating a model is very difficult. Creating a model is possible only if there are significantly more complete entries in the set. There are two ways one can approach this problem. One way is by finding the causes of missing values and removing them, and another way is using the missing values as a quasi-variable.

Removing the causes of missing values is a quick and efficient way to create more usable data. Missing values are often caused by certain columns having very little recorded values, but can also be due to different columns representing the same data. Both options are observed within this data set, so removing these causes would be helpful. By removing "problem variables" the number of complete rows will increase, and the process can stop whenever a desirable number of complete rows is achieved. This method will lead to having only variables which are present in the data and means having fewer cases to analyze, which makes modeling computationally cheaper and much faster. The biggest downside to this method is that data is removed from the set, and this is data that would otherwise be extremely significant.

A second way to approach the problem is by creating a quasi-variable for the missing values in the data set. This can be done in one of two ways. Since every variable in the data set is categorical, and most are binary, each NA or missing value can simply be replaced with a dummy value or category. By doing this, the entire data set is forced to become complete. This has the upside of keeping all information, but also leads to having a large number of variables. Additionally, since most of the variables are binary, adding a third category could lead to computation times growing exponentially. Given the major time commitment of the second option, the option of removing the problem will be explored first, and if it proves to be detrimental, then the second method will be explored.

5.1 Removing "Problem Variables"

Carefully choosing which variables to remove will help ensure understanding of what information is being lost, and that will be accounted for that in the analysis of the final model. There are two ways in which variables (columns) will be omitted from the data set. The first way is by analyzing what the variable represents. By understanding what the variable means, one can logically conclude if it will be helpful, or if it should be removed. A second approach is to quantify the amount of incomplete entries that are caused by the variable, or a group of variables. By looking at missing values, one can see what variables are missing in tandem with each other. This would help us to better understand the structure of the data set, and provide a better understanding of what causes the missing values to appear.

5.1.1 Logically Removing Variables

There are two variables that can be identified in the data set that meet this criteria. These variables are ”Data Source” and ”Year Of Registration.” Both of these variables are representative of the manner and time in which a case was added to the data set. Data source is whether the case was reported over a hot line managed by IOM, or through a case manager on the victim’s behalf. The year of registration is the year in which a case was added to the data set. Since these two variables only describe the reporting process, they will not be helpful in the process of identifying victims within a country, and can be removed without having a negative impact on the effectiveness of any models.

A second type of variable that can be removed are those which serve to summarize other data contained within the data set. There are a few examples of variables which are concatenated versions of other variables and provide a written text summary. These variables are:

- Means of Control Concatenated
- Type of Exploit Concatenated
- Type of Labour Concatenated
- Recruiter Relationship

In a similar category as the previous variables, the ”Majority Status” variable serves to identify whether or not a victim was an adult at the time they were exploited. However, the ”Age Broad” variable already covers age information, and including ”Majority Status” would essentially serve as a summary variable of the age information. While it is true that the age of majority is different in various countries, the Age Broad variable is a more specific representation of the characteristics of the victim.

As a result of these findings, the following variables will be removed:

- Data Source
- Year Of Registration
- Means of Control Concatenated
- Type of Exploit Concatenated
- Type of Labour Concatenated
- Recruiter Relationship
- Majority Status
- Majority Status at Exploit
- Majority Entry

After removing these columns, our data set now has 798 complete cases. This is an improvement, but it is certainly not enough to make any meaningful model. However, quantitative methods will yield better results.

5.1.2 Quantitatively Removing Variables

The second way that individual variables can be removed is by using quantitative methods. There are many processes to complete this task, and a handful of them will be applied to this data set.

One way is to simply look at all the different values that the variable takes. If it is observed that all the entries for a variable in the data set are either NA or 1, then it is clear that any complete row will have a value of 1 for that variable. This means that the model will only take in the value of 1 for that feature in each row. Thus the variable will have a null effect on the model. This process led to the removal of:

- Is Forced Military
- Is Organ Removal
- Type of Labour Mining/Drilling

After removing these variables, there are still only 798 complete entries. However, the removal of these variables can do nothing but help us, as there is no way they can have an effect on our model. Unfortunately, these variables are the only type that can be removed and have no negative consequences. Any other variables which are omitted will have downsides, and it would be beneficial to try to find which variables are having a significant effect on the number of missing values, and to remove those. Since each absent entry is given a value of NA, a count can be conducted on the number of NA's in each column, to get a sense of which ones are contributing the most to the lack of complete entries.

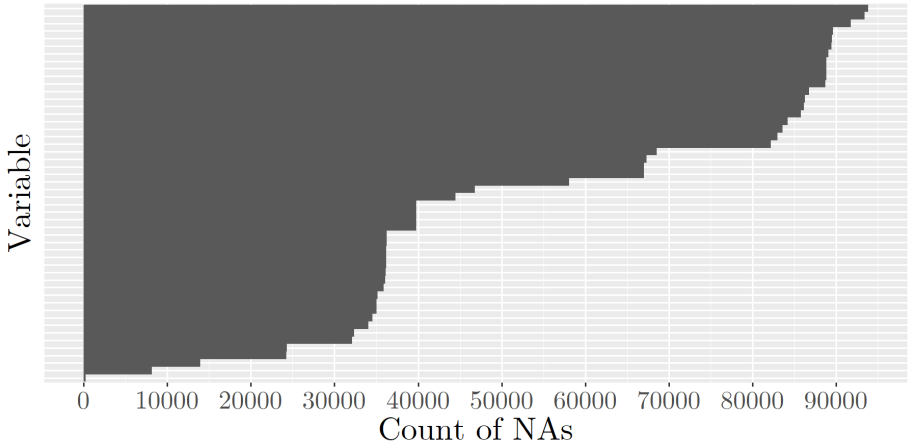


Figure 5: Count of missing entries for each variable in data set, sorted by total count

Given the large number of variables, the variable names have been omitted from the visual. However, some important information can still be gathered. One can see that every variable has at least a handful of missing values. As such, it would be helpful to start with the variables that have the highest count, and see if there are any patterns. There are no variables missing between 70 and 80 thousand entries, but there are large number of variables missing more than 80,000 entries. This led us to believe that it may be useful to analyze the variables missing more than 80,000 entries specifically.

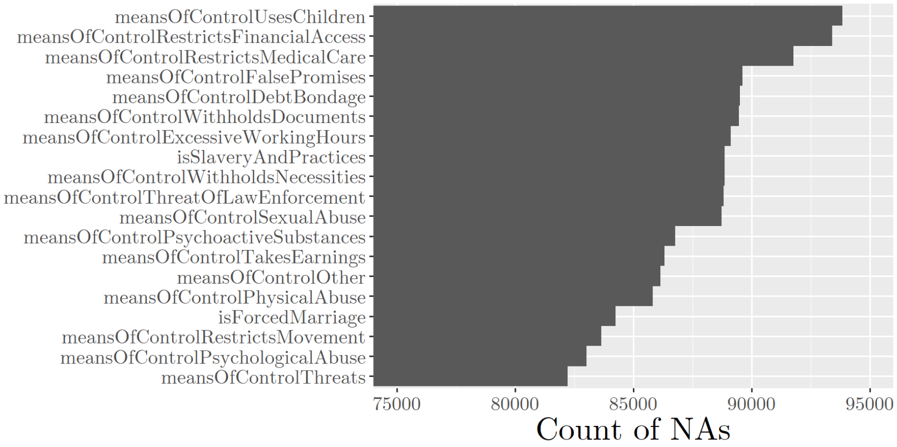


Figure 6: Count of missing entries for variables missing at least 80,000 entries

From the figure, one can see that the ”meansOfControl-” variables take up a large number of spaces on the list. Of the 19 variables with over 80,000 NA values, 17 of them are ”meansOfControl-.” This could be a direct effect of the way in which the data is recorded. If an individual is transcribing cases to the data set, they may have decided that after determining that one type of control was used, to leave all the other types as ”missing.” The data set does have a variable that is 1 if there is no specified means of control. This variable was further analyzed to determine if it would be worthwhile to pre-process the data set in a way that salvages the means of control information that does exist.

The ”means of control not specified” variable takes a value of 1 if there is no specified means of control. This variable has roughly 50,000 values of 1, and roughly 30,000 values of 0. Meaning that only 30,000 entries in our data set have a means of control specified. By replacing each NA in these variables with 0, then by adding up all the values for each entry in the data set, this will tell us how many means of control variables are specified.

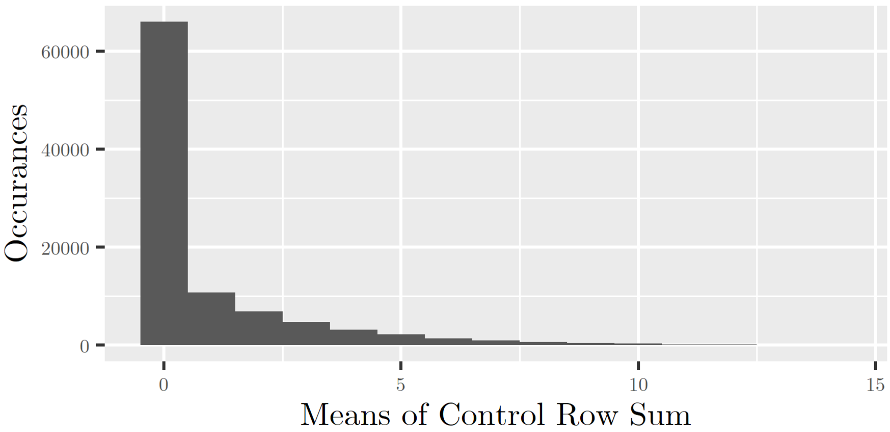


Figure 7: Number of rows in which a certain number of meansOfControl are observed.

This histogram allows a visualization of how many instances there are of means of control not being specified. From this histogram, one can see that row sums of 0 are the most common and make up over two-thirds of the data entries. This means that a large proportion of our data cannot even be salvaged by replacing NAs with values of 0. Even if one were to choose to replace the NAs with 0, the sheer lack of recorded entries implies that this data is hard for law enforcement to notice. As such, even if these variables were considered to have a significant effect on model outcomes, there are so many of them that the model would place a higher importance on these types of variables. This is a result of over fitting, and since the goal is for the final model to be applicable and easy to understand, removing these would lead to that goal the quickest.

After removing the Means of Control variables, there are 32 variables, and 806 complete entries. This is great progress, but more can be made. To get a better idea of where missing values are, the following visualization is helpful:

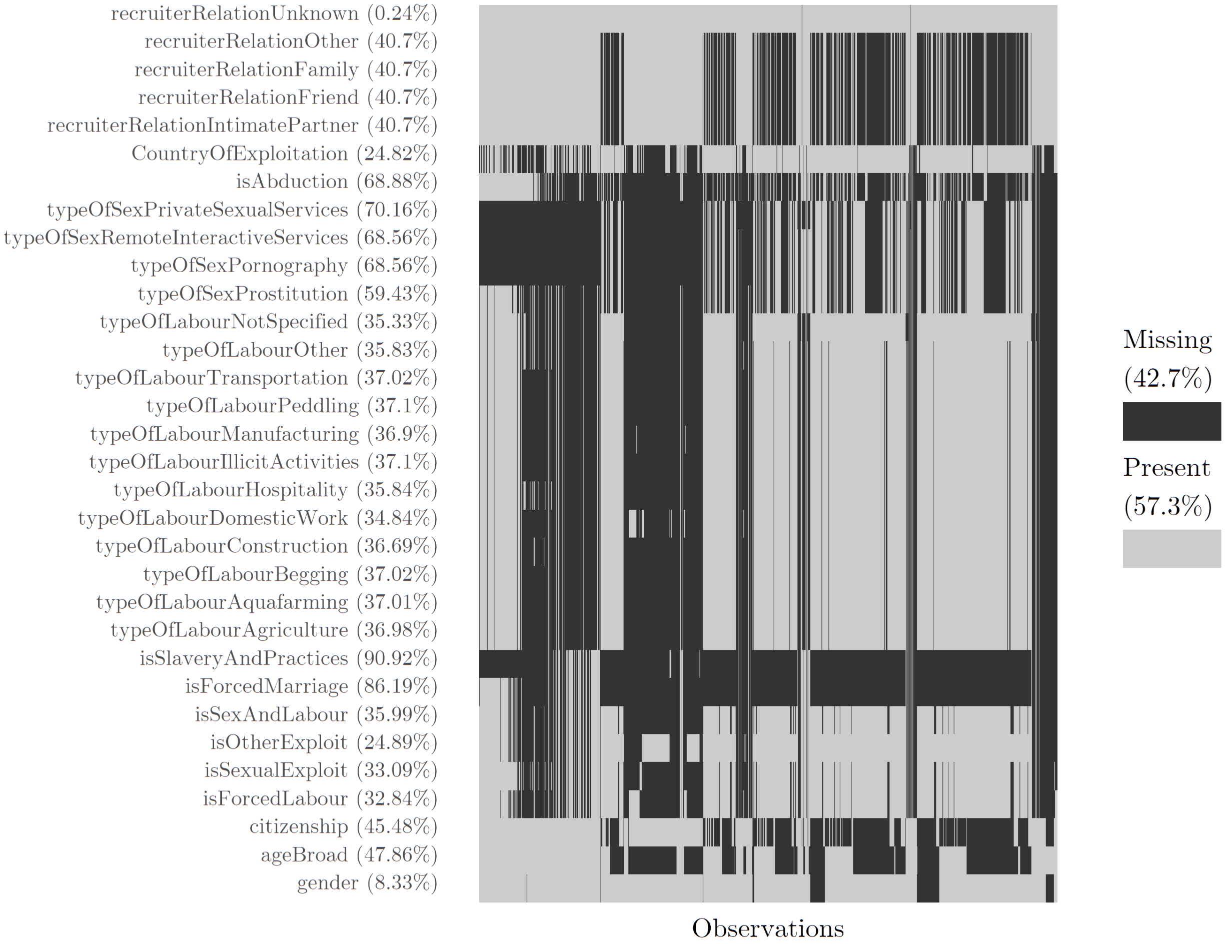


Figure 8: This is a visual of every missing value in the data set. Each row represents a variable and each column represents a case of trafficking. Each black rectangle shows where a missing value is located within the data set

In this visual, the missing values for each variable are shown visually. One can see large bands of missing data in both the horizontal and vertical directions. Horizontal bands of missing data correspond to that variable having a lot of missing values. For example, isSlaveryAndPractices is missing 90.92% of its values, and as such that row seems to be almost entirely black.

Additionally, there are instances when there seems to be a large correlation between values missing in different variables. An example of this would be the ”recruiter relation” variables. In virtually all cases in which one of these variables is missing, all the others are missing as well. This is interesting, as it would imply that if the recruiter relation was known, values of 0 were entered for the recruiter relationships that were not present. This is contrast to the means of control variables, in which NA was evtered even if another means of control was observed. This leads one to believe that there is inconsistent data entry practices within the IOM.

From this visual, there was evidence to remove "isSlaveryAndPractices", and "isForcedMarriage", as they have percentages of missing values, 90% and 86% respectively, that are significantly higher than the other variables. Removing these variables brings the number of complete entries from 806 to 1,424 complete cases. Removing these two variables has more than doubled our number of complete entries.

While 1,424 is enough complete entries to create some meaningful models, it is still only representative of less than 2% of all the entries in the data set. It would be a good idea to look at some interactions between missing variables, to see if there are any that are missing together as opposed to individually to yield better model results. This can be done with an Upset Plot, as is outlined in Lex and Gehlenborg 2014.

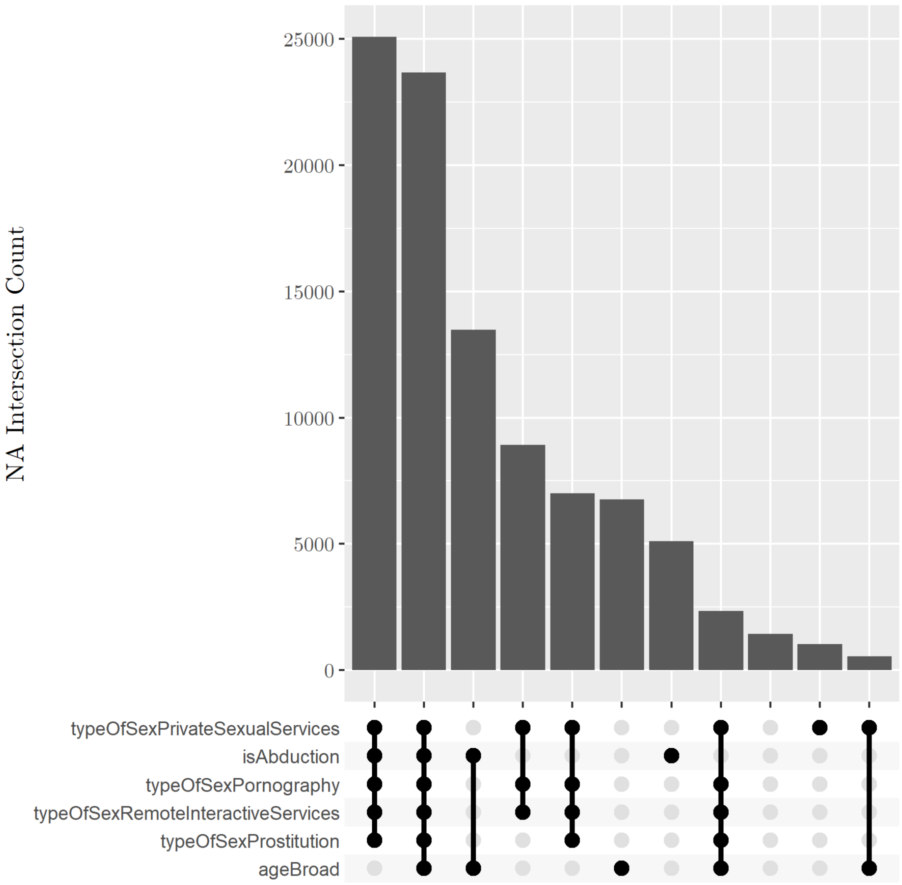


Figure 9: Count of cases in which every variable selected under the chart is missing

From this visualization, along the bottom, one can see which sets of variables lead to the most missing values. For example, the ”type of sex” variables and ”isAbduction” have a count of roughly 25,000. This means that there are 25,000 rows in our data set that are missing all of those values. This gives us a visual way to determine which variables contribute to incomplete entries, and give us the exact number of rows that are incomplete due to those combinations of variables. Thus, one can attempt to find groups of variables that have a high number of missing rows, but that are also ineffective metrics that law enforcement and policy makers can use.

The first set of variables have over 25,000 incomplete rows; however, removing 4 variables seems like it would be too much, especially since tis group includes ”isProstitution” and ”isAbduction,” both of which are variables that have been shown to be correlated with high rates of human trafficking (Tiano

2012; Polaris 2017). The first set that does not contain either of these two variables is set 4, which contains private sexual services, pornography, and remote interactive services. These are variables that one can feel comfortable removing, as they contribute to a large number of incomplete rows, and are hard to find and prevent from a law enforcement and policy perspective.

Removing these variables yields 6,672 complete cases with 27 total variables. While this process did not take into account the root causes for missing data, later discussion will attempt to determine these causes, and explain what effect it has on the final model. At this point, there are a large enough number of complete entries with a small count of variables, enough for the modeling process to begin.

5.1.3 Final Removal

After all the changes made, there are now more colomns in which every value is the same. Naturally, these columns will be removed from the data set, but the number of complete entries will remain he same. These variables are:

- Type Of Labour Illicit Activities
- Type Of Labour Peddling
- Type Of Labour Transportation
- Type Of Labour Not Specified
- Is Abduction

The previous methods of data augmentation have allowed us to have a complete data set; however, some unintended consequences of these steps have ultimately occurred. These consequences are not the result of any individual step, but are due to a combination of factors. The most significant contributing factor is perhaps the non-uniform distribution of missing values. This fact is best illustrated when the stages of demographic transision are added to the data set.

5.1.4 Augmenting Demographic Transition Data

As previously mentioned, a country's stage in demographic transition can be determined by looking at a current population pyramid for that country. To accomplish this, the United States Census Bureau has publicly available data of every country (*International Database* 2023), including current population pyramids. By observing these, one can determine a country's current stage in demographic transition. This information was then appended to the CTDC Data Set and will serve as the dependent variable in this research. These values will replace country names in the "citizenship" and "CountryOfExploitation" variables.

5.1.5 Correlation

Another metric that will help determine the suitability of the data set is by observing the correlation matrix of all variables. In this matrix, darker squares indicate a stronger correlation between two variables. Places in which the squares are lightest indicate little or no correlation. Additionally, after a hypothesis test is completed, an × is placed over any combination of two variables with an insignificant correlation (significance level of $p <= 0.05$). This means that any two variables without an × have a correlation that is assumed to be significantly different from 0.

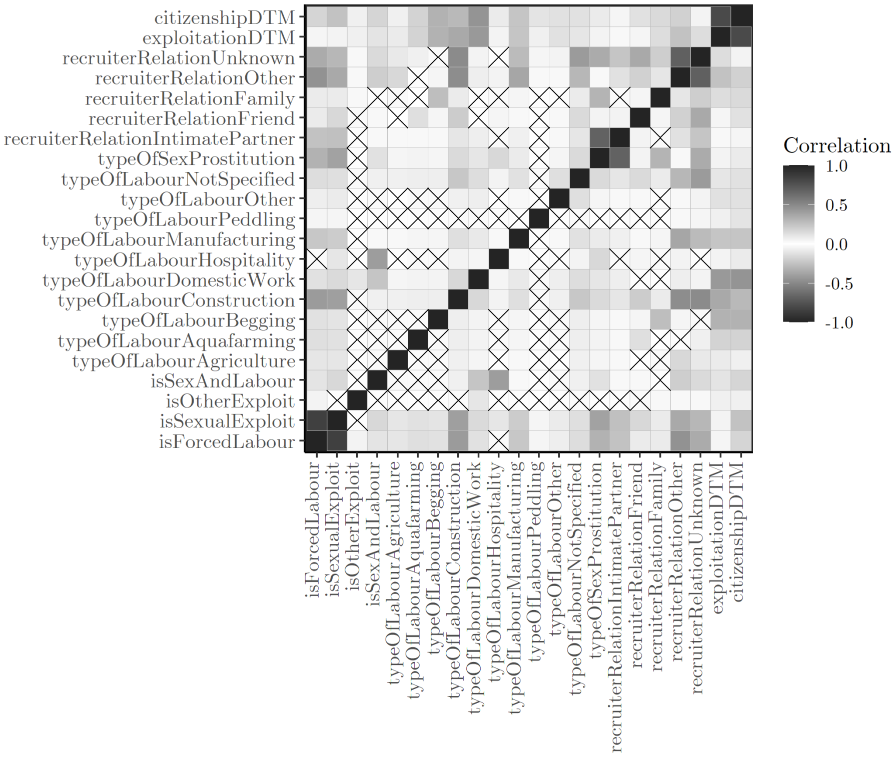


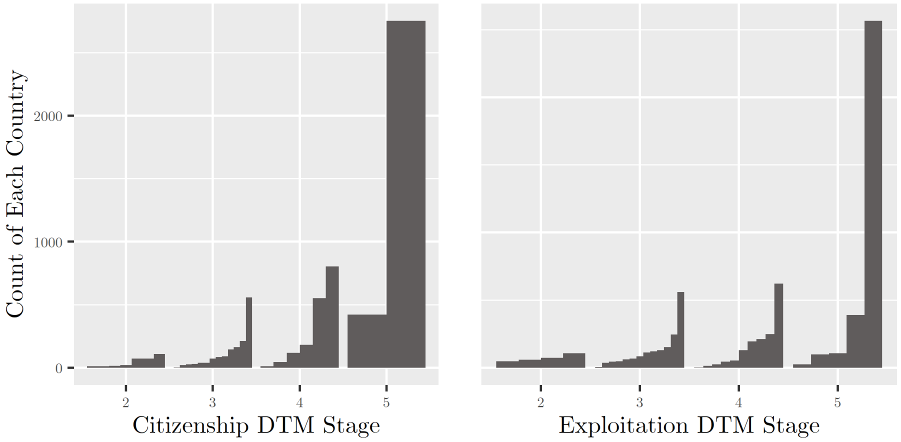
Figure 10: Correlation Plot Of All Variables

From the figure, there are no instances of a variable having insignificant correlation with all the other variables. Additionally, there are no observed variables which have a high correlation with all other variables. This is a good sign, because it means there will not likely be any problems of collinearity. Another interesting observation is the fact that citizenshipDTM and exploitationDTM appear to have comparable correlations with each variable in the data set. However, this is to be expected, as it has been shown that cases tend to have the same value for both variables.

From this plot, it is shown that the exploitatioDTM of a country is most correlated with typeOfLabourDomesticWork and typeOfLabourConstruction. The same is true for citizenshipDTM, and since there are no variables that are insignificantly correlated with all other variables, there is no need to make any major augmentations to the data set. This means that the data set could be a contender for creating thorough and significant models.

5.1.6 Individual Country Over-representation

Since the data set has only 4 categories for DTM stage, there needs to be assurance that none of the stages are being over represented by one specific country. If all of our stage 2 entries originate from one country, then there can be no conclusion that the model will generalize for all countries, and this is where a major problem of the previous steps emerges.



In these bar plots, each bar corresponds to an individual country, and the height of the bar shows the count of that country in the data set. In both cases, there seems to be no countries that overpower stages 2, 3, & 4. However, in both graphs, there seems to be one country in stage 5 that has significantly more entries than the other stage 5 countries. In order to determine if this will effect any modeling, one would need to analyze the data of the countrys in question, and determine if there is enough variation in the data to still use it.

From further analysis, it is determined that the stage 5 citizenship country with a high count is UA (Ukraine), and the high count in the stage 5 exploitation countries corresponds to RU (Russia). Since these two countries have a history of conflict, future research may yield interesting findings, given that the data set has so many entries. Additional modeling could be completed purely on Ukranian citizens who are exploited in Russia. However, this has been a long known phenomenon, in which Ukrainian citizens are exploited in Russia. This data set predates the 2022 war between Ukraine and Russia, but it has been noted that the Russian government likely has some involvement in the trafficking of persons, hence why it is so widespread in the country (*Human Trafficking: The Secret to Putin's Economy* 2020).

While further analysis of Ukraine and Russia would be interesting, the overall goal is to create a general model for all countries. To determine if the effect of these two countries is too strong, there needs to be a determination if there is a significant difference between the overall data structure between these countries, and other stage 5 countries. This can be achieved with a Chi-Square test.

For the sake of the Chi-Square test, the data set will be split into two groups. Group A will be all entries which are the over represented country, and group B will be all other countries that are in the same stage. A Chi-Square test can look at each variable, and determine if there is a significant difference between the two groups for that variable. Since one group is exclusively made up of one country, this will show if that one country has a significant difference between other countries of the same stage. From this, ideally there would be no significant difference between the groups in regards to any variable.

However, the Multiple Comparisons Fallacy is likely to be present. This means that for any large group of data, if a significance test is completed over every variable, eventually a significant P-Value will be found (or insignificant depending on the case). What this means for the following test is that we are hoping to find many instances of there being no significant difference between the groups, but there may be one or two insignificant p-values, and this does not discredit the entire analysis. This is further explained in Goodman 1999.

For the following Chi-Square tests, the null hypothesis is that group A and B not significantly different, and the alternative is that they have do have significant differences for that variable. That means that a significant p-value means that there is a significant difference between the groups. Additionally, not every variable will appear in the table, those that are omitted have the same value for all entries, and thus a test would not be possible, and would not give any meaningful information.

Table 2: χ^2 Comparing Russia to Other Stage 5 Countries of Exploitation

Variable Name	χ^2 **Statistic**	p	$p < 0.05$
gender	609.9	1.177×10^{-134}	$*$
ageBroad	482.9	3.804×10^{-100}	$*$
isForcedLabour	621.4	3.801×10^{-137}	$*$
isSexualExploit	621.4	3.8014×10^{-137}	$*$
typeOfLabourAgriculture	14.5	1.405×10^{-4}	$*$
typeOfLabourConstruction	297.1	1.443×10^{-66}	$*$
typeOfLabourDomesticWork	61.5	4.507×10^{-15}	$*$
typeOfLabourHospitality	57.3	3.750×10^{-14}	$*$
typeOfLabourManufacturing	2.9	0.085	
typeOfLabourOther	4.3	0.037	$*$
typeOfLabourNotSpecified	35.5	2.595×10^{-9}	$*$
recruiterRelationFriend	68.5	1.280×10^{-16}	$*$
recruiterRelationFamily	0.1	0.724	
recruiterRelationOther	131.4	2.050×10^{-30}	$*$
recruiterRelationUnknown	265.4	1.147×10^{-59}	$*$

Table 3: χ^2 Comparing Ukraine to Other Stage 5 Countries of Citizenship

Variable Name	χ^2 **Statistic**	p	$p < 0.05$
gender	390.4	1.674×10^{-85}	$*$
ageBroad	950.4	6.812×10^{-194}	$*$
isForcedLabour	695.4	9.878×10^{-152}	$*$
isSexualExploit	695.4	9.878×10^{-152}	$*$
typeOfLabourAgriculture	11.6	2.961×10^{-3}	$*$
typeOfLabourConstruction	229.2	1.725×10^{-50}	$*$
typeOfLabourDomesticWork	50.7	9.973×10^{-12}	$*$
typeOfLabourHospitality	35.1	2.4445×10^{-8}	$*$
typeOfLabourManufacturing	81.9	1.652×10^{-18}	$*$
typeOfLabourOther	4.2	0.121	
typeOfLabourNotSpecified	29.3	4.416×10^{-7}	$*$
recruiterRelationFriend	30.6	2.267×10^{-7}	$*$
recruiterRelationFamily	1.4	0.487	
recruiterRelationOthe	340.9	9.269×10^{-75}	$*$
recruiterRelationUnknown	435.1	3.345×10^{-95}	$*$

From the test for stage 5 exploitation countries (Table 1), one can see that there is a significant difference between the two groups for 13 out of the 15 variables being tested. This means that there is a significant difference between Russia and other stage 5 exploitation countries. This means there is an over representation problem, in which one subset of stage 5 countries (Russia) is having a significant effect on our overall data set. This is comparable to sampling bias, and similar measures can be used to combat the issue. Also note that these results are only a problem due to the large number of entries. If it was determined that the countries were not significantly different, then no action would need to be taken. It is due to the large number of entries and the lack of similarities that are the root of the issue, neither one of these facts is an issue on its own.

Similarly, in Table 2 there exists the same problem with Ukraine, despite there being more stage five countries represented in citizenship than with country of exploitation. Since these are both instances of oversampling from one population (Ukrainian citizens and Russian exploits), this can be counteracted by applying a weight to these instances in the data set. Another solution would be to randomly select entries in which Russia is the country of exploitation or Ukraine in the same of citizenship. This allows us to bring the raw sample count down so that these countries are not over represented.

The process of weighting means that instead of letting these countries (Russia and Ukraine) have the same effect on the model as other countries in their respective categories of exploitation and citizenship, one can assign a weight to these cases. This weight would be less than 1, since we want these cases to have a smaller effect on the model. This process is outlined in Wirth and Tchetgen 2014, and would be implemented in the modeling phase.

5.1.7 Conclusions From Removing Causes of Missing Data

Initially, the act of creating complete entries was done so in an effort to reduce computational time, simplify data analysis, and allow simple model creation. However, in doing so, a combination of factors has led to the creation of a biased data set. Looking again at figure 8, one can see that the distribution of missing values is extremely non-uniform, implying that there may be a significant effect that other variables have on a variable being missing. For example, it could be that every case located in a particular country may have the same value missing. Additionally, there are some cases in which most values are missing. This could be cause for concern, but is not an insurmountable issue.

The second major problem is the over-representation of Russia and Ukraine, this could be due to having a larger number of cases than other stage 5 countries, or it could also be indicative of data reported from these countries containing more information and fewer missing values. Either way, the oversampling of these two countries is extremely significant, and is cause for concern. While it would be possible to solve this problem by selecting a sub-sample of these entries for our data, this would greatly reduce the number of usable entries, and would lead to ineffective and insignificant models. It is for these reasons that the process ofcreating a data set that implements ”missing-ness” as a variable, in hopes of creating a larger data set.

5.2 Quasi-Variable Representation of Missing Values

To start, the initial large data set will be revisited. With 63 columns and 0 complete entries, we will now begin the process of creating a second potential data set for modeling. This data set will attempt to implement information gained from knowing which variables are missing, instead of ignoring missing data entirely.

5.2.1 Salvaging Missing Data

To start, the summary variables will be removed, for the same reasons as mentioned previously. These variables serve as a text summary of other attributes of a case which are already recorded as binary data. These removed variables are:

- Means of Control Concatenated
- Type of Exploit Concatenated
- Type of Labour Concatenated
- Recruiter Relationship

After removing these variables, each missing value (NA) in the data set is assigned a value of -1. This allows us to have a simple way to notate and keep track of which values are missing. A value of -1 is not a possible value anywhere in the data set, so it can be safely used to denote when a value is missing. A good objective would be to replace some of these missing values with values of 0 or 1. To begin this, it was realized that there were many instances of missing variables in a variable category (such as type of labour), even when other values were recorded as 1 in that category. It is believed that this is a result of individuals recording data into the database only putting values of 1, but not values of 0 into the data set. It is impossible to know exactly how the data was recorded, but simple assumptions such as these must be made in order to gather useful information which would otherwise not be usable.

There are many instances in the data set in which a value of 1 is recorded for a sub-variable within certain variable categories, and all other variables in the category are missing. It is reasonable to assume that this is potentially a mistake, and these missings should actually be zeros. So each row in the data set was checked, and instances such as these were "salvaged" by replacing the missing values with 0. The following table outlines the number of rows that were salvaged in each category of variables.

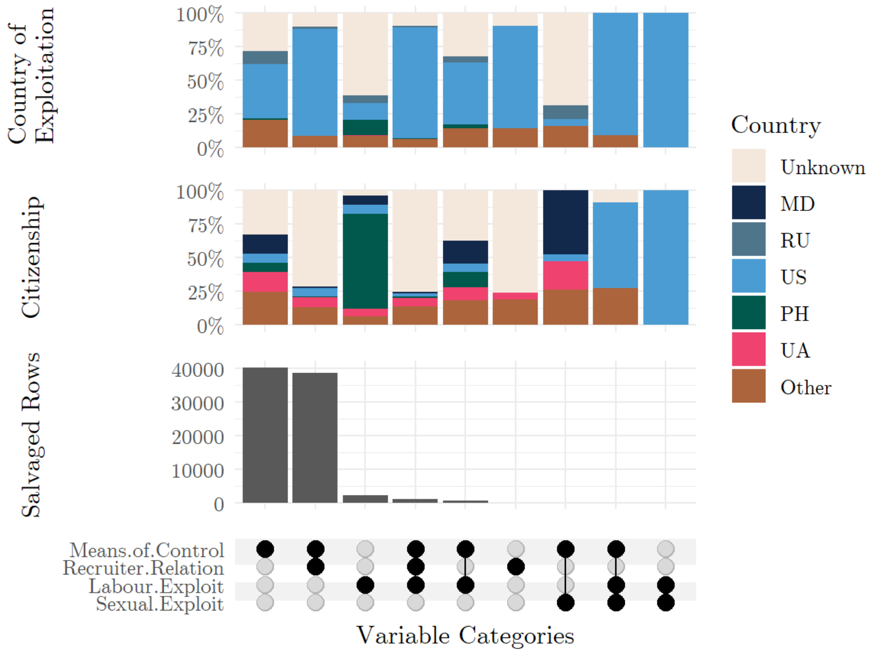
Table 4: Number of Salvaged Rows By Category

Variable Category	Number of Salvaged Rows
Sexual Exploit Type	31
Labour Exploit Type	3,979
Means of Control	80,325
Recruiter Relation	39,561

One can see from the table that there were many instances in the data set in which one sub category or type was observed, and others were left as missing. this is understandable as it could be a truly unknown value, but since the goal is to inform law enforcement on what they should be looking for, if a value is unknown when the victim is saved, it is logical to conclude it could not have been observed by law enforcement. However, it is still important to gain some more insight into where the missing values occur.

First, it is good to visualize which rows were salvaged in the table, and seeing if these correspond to particular countries.

Figure 11: Types of Rows Salvaged By Country



To break down this figure, we start with the bottom section that consists of circles. On the left of this section is a variable category such as Labour Exploit and Recruiter Relation. As previously mentioned, all of these categories include binary variables for different aspects of the category. Aspects such as if the labour exploit was in agriculture, and if the recruiter was a family member. In each column of the plot, the dark circles indicate which combination of these categories were salvaged. For example, column 1 corresponds to rows in which Means.Of.Control was the only salvaged category, and column 2 corresponds to rows in which Means of Control and Recruiter relation were both salvaged.

The next section, ”Salvaged Rows” simply provides a bar plot that counts the number of rows that were salvaged for each category combination. Finally, the Citizenship and Country of Exploitation sections give a percentage of the salvaged rows that were made up by certain countries. Only the 5 most common countries that make up the salvaged rows are noted, with all other countries being represented by "Other" and NA values as "Unknown".

From the visual, we also see that the most frequently salvaged rows are the ones involving Means of Control and Recruiter relation. This means that these variables were initially underrepresented in the data. This could be due to the fact means of control and recruiter relation are often overlooked in the realm of investigatory knowledge and victim assistance. This is not an unexpected result, as it is often the case in the United States and many other countries, that the manner in which a victim is exploited is perceived as more important than what led to the victim being exploited to begin with. This phenomenon is further explored in Wallinger 2010.

Some surprising information gathered from this visual is the fact that the United States (US) is the largest contributing exploitation country to these instances of salvageable row. As a reminder, a row can only be salvaged if a value of 1 is recorded for a sub variable, and there are other variables that have missing values. This implies that the United States records data on victims in such a way that if an aspect of trafficking is not observed, it is not immediately assigned a value of 0. This may appear to be an issue of "sloppy" or inefficient record keeping, but it is also likely that if an aspect of trafficking is not observed, there is no way to truly know if it happened. For example, if a victim is rescued and it was observed that the victim was labor trafficked in an agricultural setting, there is truly no way to rule out if the victim was also trafficked in a construction setting.

This reasoning seems fairly straightforward and reasonable; it eliminates instances of type II errors (false negatives) as a result of difficulty investigating trafficking. These errors could even occur as a result of a victim's inability to communicate or even remember their experiences, which is a frequent occurrence among victims of trafficking (Oram et al. 2016). Essentially, it is extremely difficult to determine that a certain aspect of human trafficking did not happen; additionally, from a policy and enforcement perspective, it does not make sense to make claims and decisions based solely off of an aspect of trafficking being recorded as 0 instead of NA.

By minimizing type II errors, policy makers are able to make decisions based off of the aspects trafficking they know to exist in their jurisdiction. These same policy makers would also be unable to preemptively "rule-out" certain aspects of trafficking, which could later become more easily recognized. So, while removing al potential type II errors int he data set seems extreme, there is reasoning to lead to this decision. This decision would also make model creation and interperetation immensely easier, as each variable would become binary (0 or 1). This binary data was the initial intent of the data set, but frequent missing values led to a data set with not a single row being complete. By replacing all NA values with 0 and interpreting values of 0 as "not observed" instead of "not present" this will lead to a new data set that better aligns with the scope of the project.

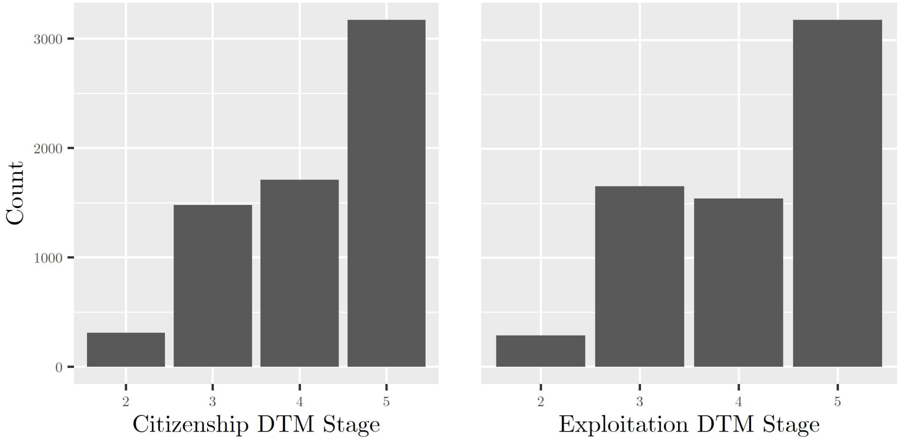
5.2.2 Representing Missing Values as 0

Like in previous sections, the summary character variables will be removed. Next, all of the intended binary variables will be updated to replace all missing values with 0. Any variables in which every value was the same were also removed. These variables were:

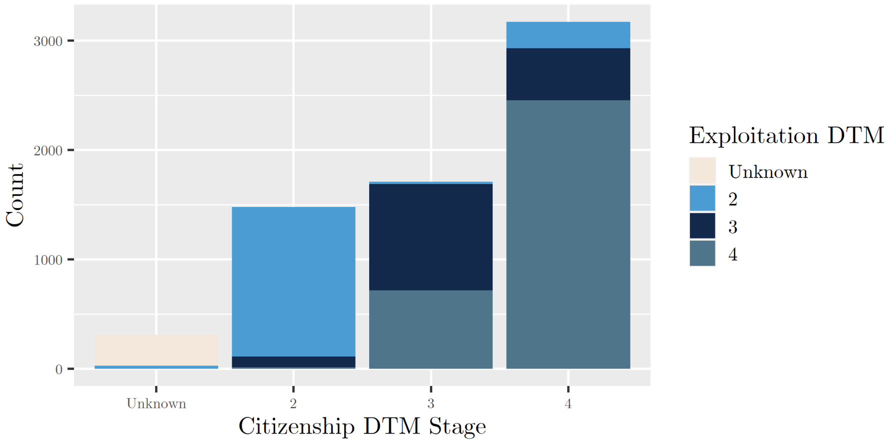
- Is Forced Military
- Is Organ Removal
- Type of Labour is Mining or Drilling

This brings our number of complete cases from 0 to 788. It was also decided to remove the "Majority Status" variables. As stated in section 5.1.1, these variables are missing most of the time, and are already represented in the broad age variable. Removing these brings our total complete entries to over 22 thousand.

5.3 DTM Stages Data

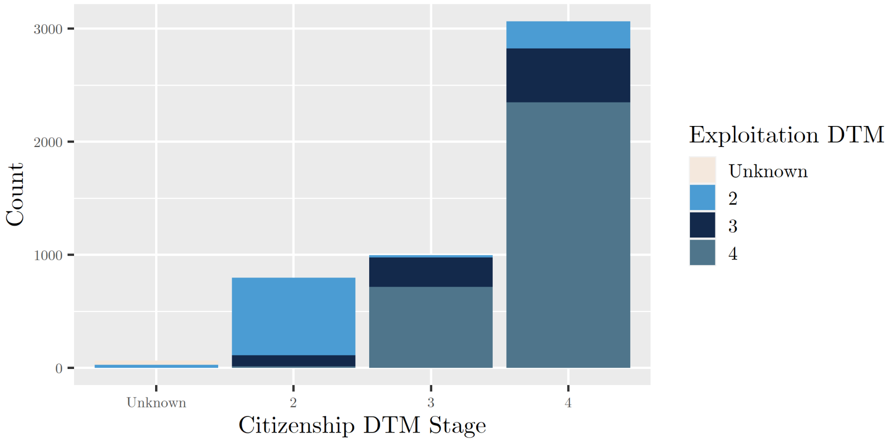


In this graphic, one can see that the majority of entries for Citizenship and Exploitation DTM are stage 5, with a decreasing count as the stage decreases. This certainly is interesting, but is something one would expect to see for the DTM stage of exploitation countries. Since these are typically the places in which the victim's case is recorded and added to the CTDC data set, this is not abnormal. One would expect more developed countries to have higher rates of reported cases, as they have access to the resources required to locate the victims. However, this does not explain the trend in Citizenship DTM Stages.



From this chart, one can see the intersection of the two plots. Each bar has the total count of entries with a certain citizenship DTM stage, and each bar is broken down into the counts of the exploitation DTM stage. For example, one can see that there are roughly 1500 victims with a citizenship DTM stage of 3, and of those, the majority are exploited in a country that is also stage 3.

Of interest in this chart is the fact that for every citizenship DTM stage, the majority of the victims are exploited in a country of the same stage. This pattern can easily be explained by victims typically being exploited in their citizenship country. Further analysis shows that 1,752 of the 6,672 entries (26.3%) are examples of this. When these entries were removed, the visualization looks like this:



This image displays a different outcome. Victims from stage 4 countries are more likeley to be exploited in a stage 5 country if they are exploited in a country that is different than their citizenship. However, stages 3 and 5 exhibit the same behavior as before, and citizens from stage 2 countries are split almost evenly between being exploited in stage 2 and 3 countries. Since thevisuals are pretty similar, one can assume that while many victims are exploited in the same country they are a citizen of, there is no evidence to suggest this will have a large impact on the outcome of our model by including these instances.

Additionally, sine the aim of these models is to inform policy makers, it would make sense to use exploitation DTM stage as the variable of choice, since these correspond to the places in which individuals are actually exploited. However, removing citizenship DTM would remove valuable information for law makers of where the victims in their country are coming from. Some potential ways to counteract this is by assigning a small weight to the citizenship DTM variable, to prevent it from overpowering the rest of the variables in the model. Additionally, adding some random variation to citizenship DTM in our models would lead to the model being more robust. This idea also would be beneficial as it would account for random variation in how DTM stage was determined.

6 Conclusion

Through exploring different ways to preprocess the data set, we have uncovered and applied different methods of handling missing data, and have thoroughly addressed the pros and cons of each method. Ultimately, the methods used for modeling will depend on the purpose and breadth of the model. We also showed some indicators of a connection between a countries stage of development and certain aspects of human trafficking. This research serves as a basis for further research which seeks to quantify these relationships, and provide a framework for policymakers and law enforcement to find ways to address human trafficking without being hindered by missing or underrepresented data.

References

Amin, Shreya (2010). "A Step Towards Modeling and Destabilizing Human Trafficking Networks Using Machine Learning Methods". In: *AAAI Publications.*

Aromaa, Kauko (2007). "Trafficking in Human Beings: Uniform Definitions for Better Measuring and for Effective Counter-Measures". In: Springer, New York, NY.

Bar, Michael and Oksana Leukhina (2010). "Demographic transition and industrial revolution: A macroeconomic investigation". In: *Review of Economic Dynamics* 13.2, pp. 424–451. ISSN: 1094-2025. DOI: https://doi.org/10.1016/j.red.2009.03.002. URL: https://www.sciencedirect.com/science/article/pii/S1094202509000210.

Bongaarts, John (2009). "Human population growth and the demographic transition". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1532. DOI: 10.1098/rstb.2009.0137.

Galor, Oded and David N. Weil (Sept. 2000). "Population, Technology, and Growth: From Malthusian Stagnation to the Demographic Transition and Beyond". In: *American Economic Review* 90.4, pp. 806–828. DOI: 10.1257/aer.90.4.806. URL: https://www.aeaweb.org/articles?id=10.1257/aer.90.4.806.

Goodman, Steven N. (Aug. 15, 1999). "Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy". In: *Annals of Internal Medicine.*

Human Trafficking: The Secret to Putin's Economy (Nov. 1, 2020). Harvard International Review. URL: https://hir.harvard.edu/putin-and-human-trafficking/.

International Database (2023). URL: https://www.census.gov/data-tools/demo/idb/#/country.

Jac-Kucharski, Alicja (2012). "The Determinants of Human Trafficking: A US Case Study". In: *International Migration.*

Kangaspunta, Kristiina (2008). *An introduction to human trafficking: Vulnerability, impact and action.* United Nations, p. 108.

Kirk, Dudley (1996). "Demographic Transition Theory". In: *Population Studies* 50.3, pp. 361–387. DOI: 10.1098/rstb.2009.0137.

Lex, Alexander and Nils Gehlenborg (July 30, 2014). "Sets and intersections". In: *Nature Methods* 11.779.

Litan, Robert E. (Mar. 1, 2000). "The "Globalization" Challenge: The U.S. Role in Shaping World Trade and Investment". In: *Brookings.*

Malakooti, Arezo (2016). "The Dynamics of Migrant Smuggling in North Africa: Focus on the Central Mediterranean Route". In: *IEMed Mediterranean Yearbook 2016.*

Omar Mahmoud, Toman and Christoph Trebesch (2010). "The economics of human trafficking and labour migration: Micro-evidence from Eastern Europe". In: *Journal of Comparative Economics* 38.2, pp. 173–188. ISSN: 0147-5967. DOI: https://doi.org/10.1016/j.jce.2010.02.001. URL: https://www.sciencedirect.com/science/article/pii/S0147596710000028.

Oram, Siân et al. (June 2016). "Human Trafficking and Health: A Survey of Male and Female Survivors in England". In: *Am J Public Health* 106.6.

Polaris (Mar. 2017). *The Typology of Modern slavery.*

Raymond, Janice G. (2002). "The New UN Trafficking Protocol." In: *Women's Studies International Forum.*

Sanchez, G. et al. (2018). "Beyond Networks, Militias and Tribes: Rethinking EU Counter-smuggling Policy and Response". In: *Euromesco Policy Study.*

Schneider, Friedrich (1994). "Measuring the Size and Development of the Shadow Economy. Can the Causes Be Found and the Obstacles Be Overcome?" In: *Essays on Economic Psychology* 90.4, pp. 193–212. DOI: 10.1007/978-3-642-48621-0_10.

Sheu, Jessica C. et al. (May 2021). "Potential Impact of Climate Change on Human Trafficking". In: *The Journal of Nervous and Mental Disease.*

The Global K-Anonymized Dataset (Dec. 14, 2021). URL: https://www.ctdatacollaborative.org/global-k-anonymized-dataset.

Tiano, Susan (2012). "Human Trafficking: A Perfect Storm of Contributing Factors". In: *Borderline Slavery.* 1st ed. Routledge.

Understanding Human Trafficking (Apr. 2022). U.S. Department of State, Office To Monitor and Combat Trafficking in Persons. URL: https://www.state.gov/what-is-trafficking-in-persons/.

United Nations General Assembly (Nov. 15, 2000). "Protocol to Prevent, Suppress and Punish Trafficking in Persons Especially Women and Children, supplementing the United Nations Convention against Transnational Organized Crime". In: *General Assembly resolution 55/25.*

Wallinger, Caroline S. (Oct. 2010). "Media Representation and Human Trafficking: How Anti-Trafficking Discourse Affects Trafficked Persons". In: *Second Annual Interdisciplinary Conference on Human Trafficking, 2010.* Ed. by editor.

Wirth, Kathleen E. and Eric J. Tchetgen Tchetgen (May 2014). "Accounting for selection bias in association studies with complex survey data". In: *Epidemiology* 25.3, pp. 44–453.