Final Report

Adam Dameron

January 17, 2023

Abstract

1 Outline Methods

- Dataset (History, purpose, etc)
- Types of models created
- Why models require complete data and fewer variables (over fitting, computing power, etc)

2 Introduce Dataset

- Summarize dataset and define variables (https://cran.r-project.org/web/packages/naniar/vignettes/getting-started-w-naniar.html)
- Explain what NA means in relation to the data set, and reasons NA may be present - Briefly describe/define missing (or complete) entries and explain why this matters
- Remind that model can only be created with the complete entries

3 Manipulating CTDC Dataset

The CTDC data set contains 63 variables and 0 of its entries are complete. This means that without any data manipulation, creating a model is impossible. Creating a model is possible only if we have a significantly larger number of complete entries. One way to manipulate the data set into having a larger number of complete entries is by removing entire columns from the data set altogether.

While this process is efficient at creating complete entries, it is important to note that we would be losing some information in the process. For example, if we remove the variable that indicates if a victim was trafficking in the mining industry, then our final model cannot incorporate that information. This is why it is important a balance is struck between having complete entries, while still maintaining as much information as possible.

Carefully choosing which variables to remove will help ensure we are aware of what information we are losing, 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, we 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, we 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.

3.1 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 590 complete cases. This is an improvement, but it is certainly not enough to make any meaningful model. However, quantitative methods will yield better results.

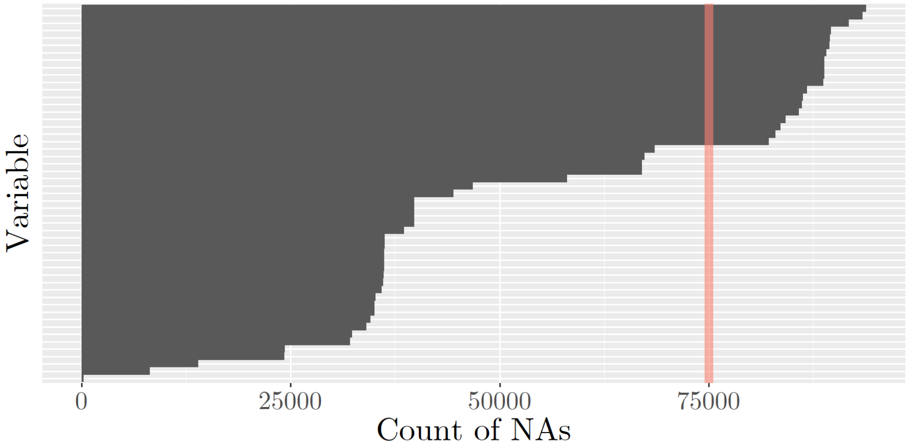
3.2 Quantitatively Removing Variables

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.

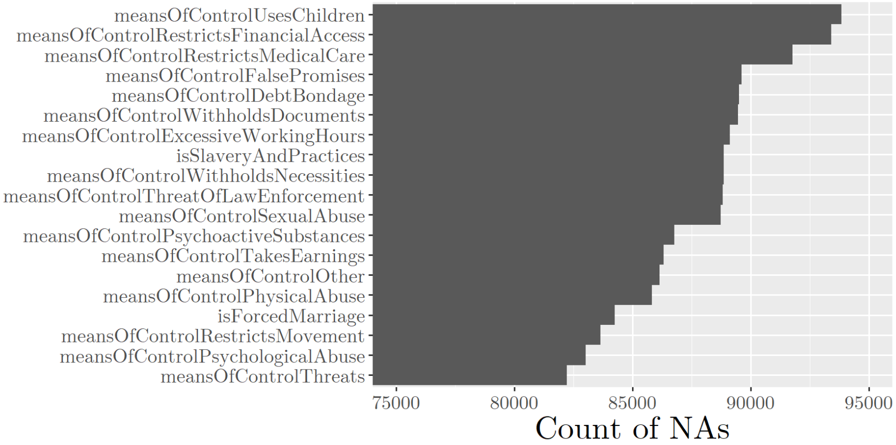
One way is to simply look at all the different values that the variable takes. If we see that all the entries for a variable in the data set are either NA or 1, then it is clear that any complete row will have a value of 1 for that variable. This means that the model will only take in the value of 1 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 590 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 we can remove and have no negative consequences. Any other variables we omit 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, we can count 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.

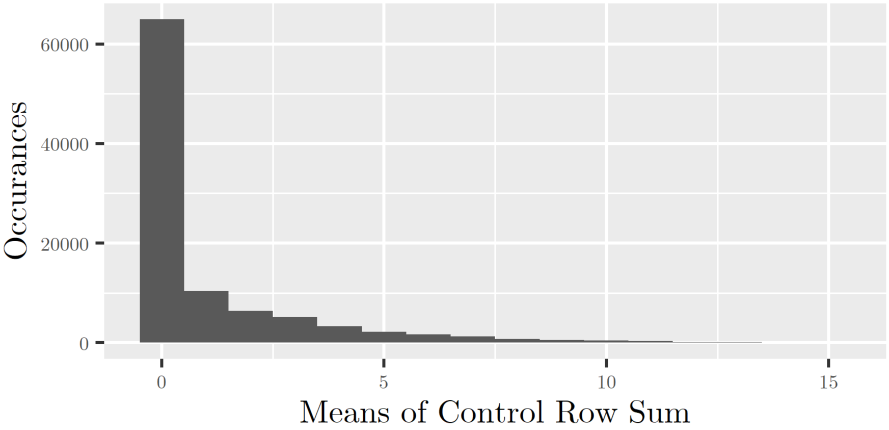


Given the large number of variables, the variable names have been omitted from the visual. However, some important information can still be gathered. One can see that every variable has at least a handful of missing values. As such, it would be helpful to start with the variables that have the highest count, and see if there are any patterns. There is a steep drop off of the count of NAs at around 75,000 (emphasized with red line), so analyzing all the variables with more than 75,000 NA could give some useful information.



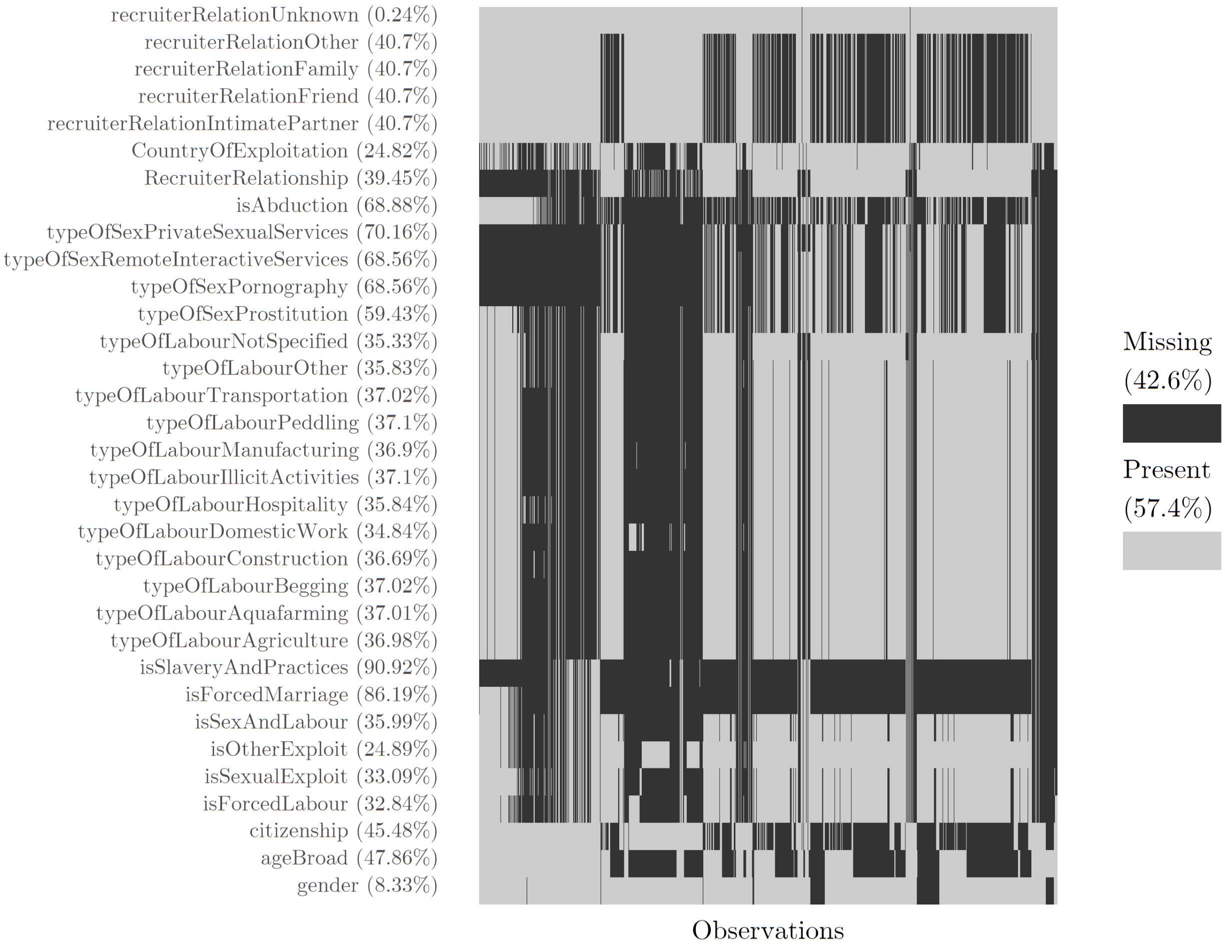
From the figure, one can see that the "meansOfControl" variables take up a large number of spaces on the list. Of the 19 variables with over 75,000 NA values, 17 of them are "meansOfControl." This could be a direct effect of the way in which the data is recorded. If an individual is transcribing cases 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. We can analyze this variable to determine if it would be worthwhile to modify the data set to salvage 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, we are able to add up all the values for each entry in the data set. This will tell us how many means of control variables are specified.



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 we want 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, we are left with 33 variables, and 596 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:



In this visual, the missing values for each variable are shown visually. Each row in the visual represents a category in the data set, and the columns show each row. For example, we 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 missings. 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 values of 0 were entered, unlike in the case of the means of control variables. This leads one to believe that there is inconsistent data entry practices within the IOM.

From this visual, we would consider removing ”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 596 to 1,218 complete cases. Removing these two variables has more than doubled our number of complete entries.

While 1,218 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. This can be done with an Upset Plot, as is outlined in Gehlenborg 2014

References

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