**Analysis of data-set. Part of LDSQ/ CAIDS-Q study. Week 22-28th Jan. Pablo M. Rodriguez.**

1. Choosing tools

The software chosen for the analysis is RStudio, a development environment for R, a programming language orientated to statistical computing and graphics. The reasons are: it is open source and free, is highly functional and, most important, allows for reproducibility. All the code is stored in a script, from reading the raw data to its cleaning, transformation, modelling and visualization, which can be executed anytime, in any computer (while having the original data located) and still the outcome (summary of data, graphs, models, etc.) will be the same. This also makes revision easier to carry out.

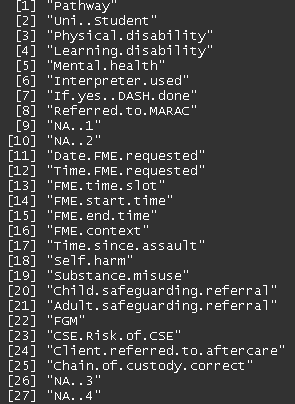
Any progress or change made is uploaded with comments into this remote repository: <https://github.com/pabrodez/LDSQ-CAIDS-Q> This repository was created to avoid losing any progress made and to enable any kind of collaboration, as any participant can download the folders on any terminal, check progress made and collaborate by uploading contributions to branches of the project. It also facilitates transparency, as anyone interested can see it but without having access to the original data, that I think we should keep not public until the study is released.

2. Progress done

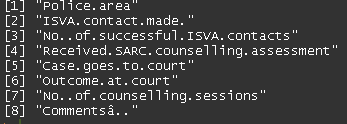
Before reading both data missing variables, empty rows and differences in denomination of same variables across both tables were looked for.

Both datasets were read into RStudio but keeping them splt into four: Adults Dec-May, adults June-Nov, children Dec-May and children Jun-Dec. Because the tables have different variables in both periods, and adults and children have slightly different ones, seems appropriate to keep them separate before binding.

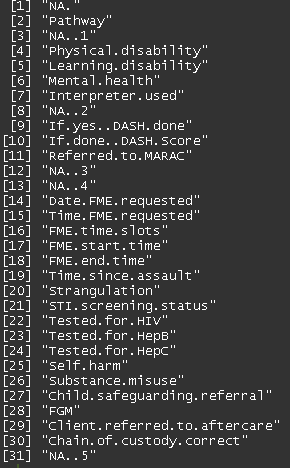
Different variables were considered for input in both periods. Variables from Jun-Nov not present in Dec-May (adults) are (both lists were generated before dropping empty columns):



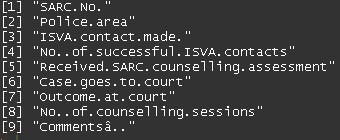
Variables from Dec-May not present in Jun-Nov are:



As for children Jun-Nov compared to Dec-May:



Viceversa:



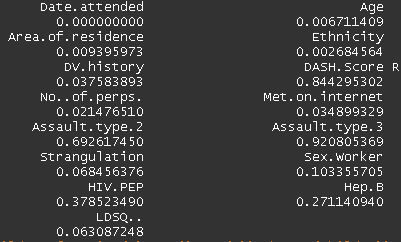
However, some variables are present in the other but named differently. It’s important to note that learning disability is a key variable to collect as it can be used to validate LDSQ scores.

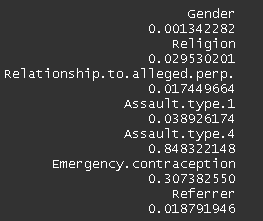
The variables "FME.start.time" and "FME.end.time" were used to compute the minutes spent in Forensic Examination, which was collected in Dec-May. "Learning disability", "Physical.disability", "Mental.health" and others not present on both were kept aside. Same variables across tables with dissimilar names were given same denomination for the sake of consistency.

Next, columns with empty values in all rows and empty rows in all columns were dropped. Also, a search for duplicated rows was made. After, unique values in all columns were checked to look for typos or invalid answers, and changed them when appropriate or substituted with NA for consistency across tables. Most common typos were wrong capitalization of words (i.e.: “NO” or “no” for “No”, or “N/A” for “NA”).

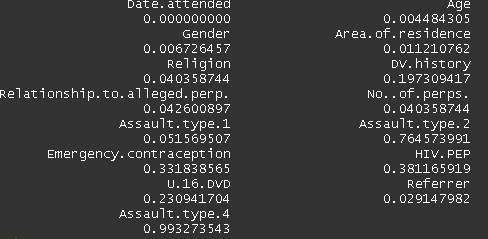
One issue that arised is the difference in mins between the start and end of FME. If, for instance, it starts at 20:00 and ends at 1:00 of the next day, computing the difference gets less easy. Because the hours do not have a date attached, the program does not know if it is the next day or the following, returning wrong results (I am working to solve this). After checking the four data-set had the same variables of the same type, they were joined into two separate table: the whole period for adults (745x22) and the same for children (446x22).

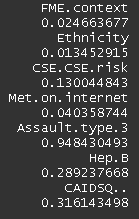
Finally, presence of missing value was explored. These figures show the percentage of NAs in the variables of the adult dataset:



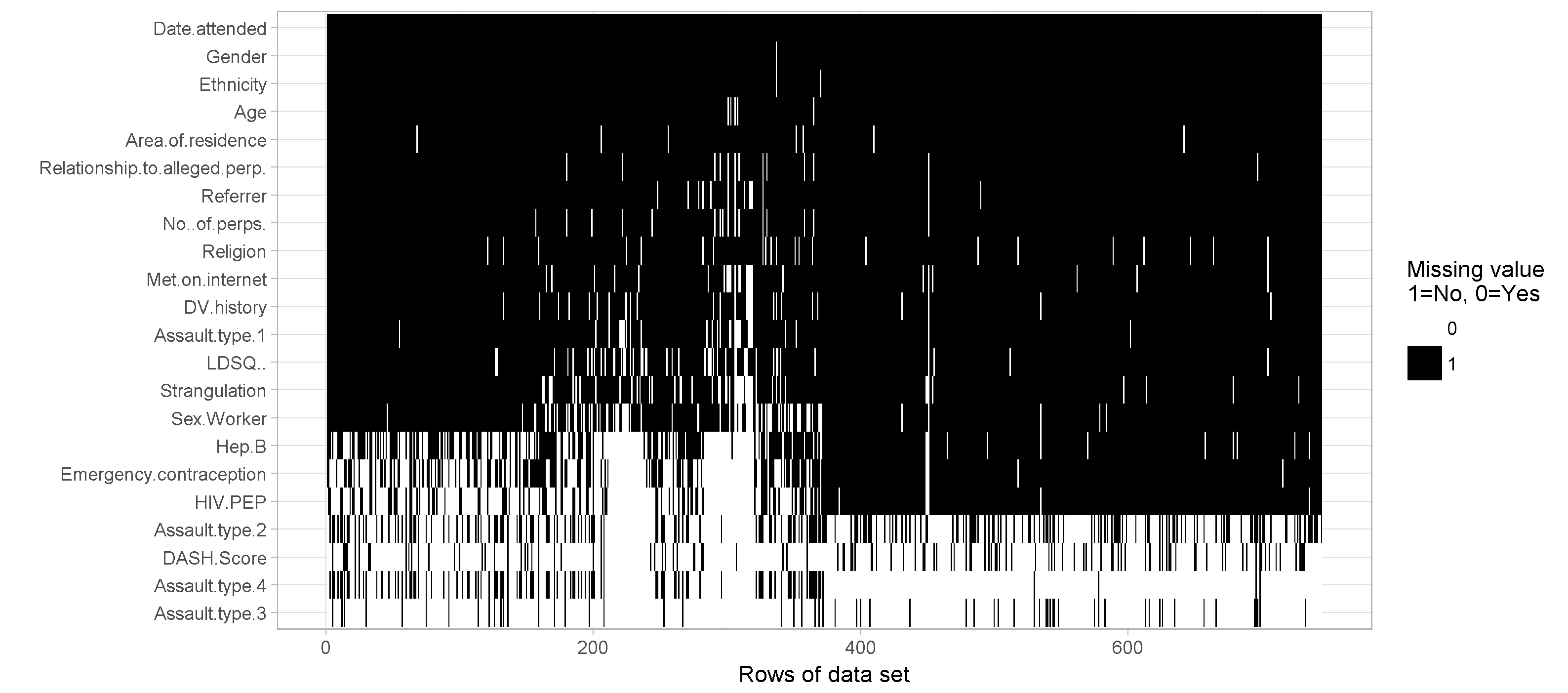


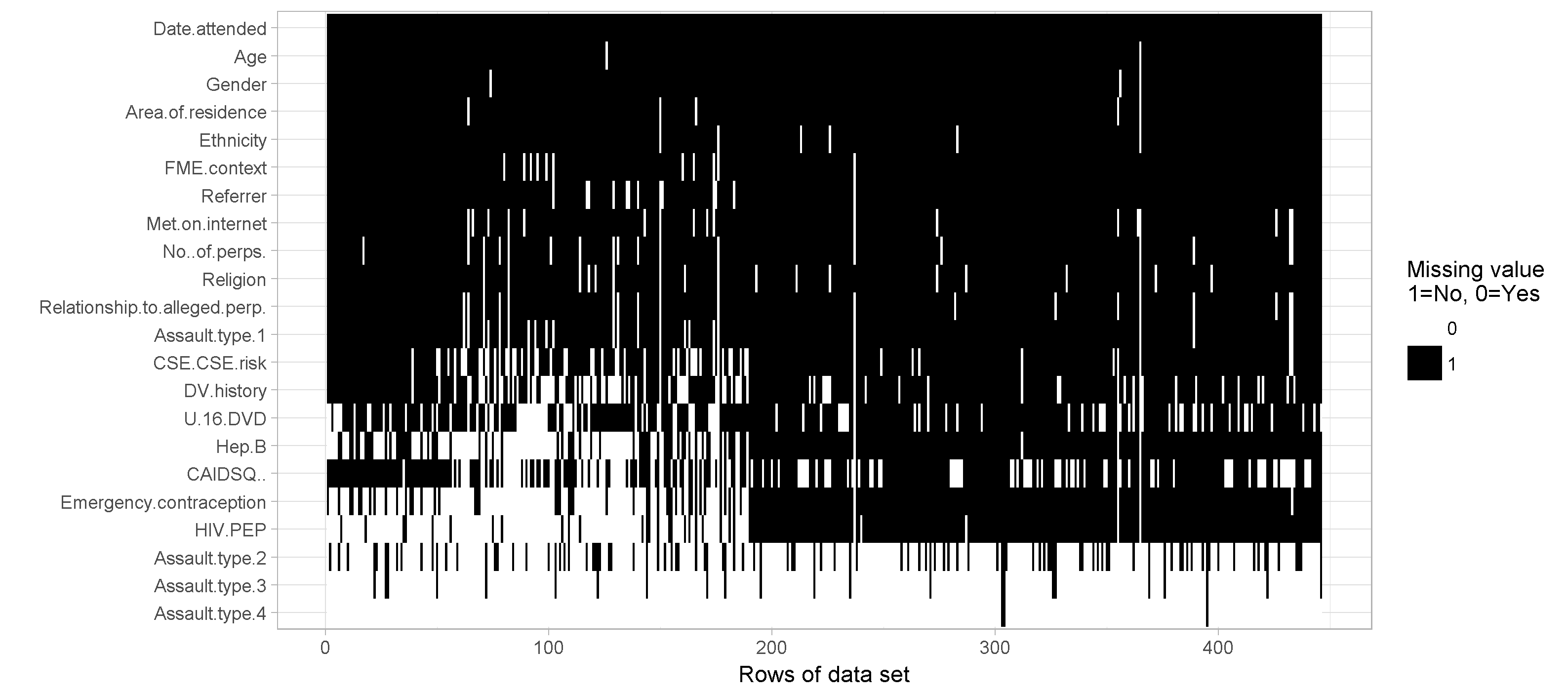
And children:





These graphs show the map of missing values across variables (adults first):





3. Next steps

- Solve the issue of difference in mins of FME start and end time.

- Review results obtained. Search for mistakes.

- Explore variables in adults and children data-sets. Carry out summary/descriptive statistics.

- Generate graphs to visualize each of them

- Explore relationships between them.