Overall, I think this data is quite limited. Most tables are useless because they are mostly empty or because their keys/IDs do not link to other tables. Some tables have some useful information, but the information they do have tends to be quite limited in one way or another. One major overarching difficulty is it’s hard to know what the *unit of analysis* is for any given table i.e., what does a row in a table represent? Does each row correspond to a specific report? Incident? Event? Unique victim? Educated guesses can be made in some cases but in other cases it’s very unclear. Without that information, it’s hard to make use of the other data we do have. For example, in some tables we can aggregate by time or by location so we could make some sort of observation about where abuse is happening or when. However, it’s unclear if we are talking about the ***number of victims*** or the ***number of reports*** or something else entirely.

I think to make the best use of this data we would really need a domain expert or someone with deep substantive knowledge to help us sift through this data.

* **Abuse types**
  + LIST\_ABUSE\_ALLEGATIONS\_DATA\_TABLE links **abuse\_type\_id** to **abuse\_type**.
  + SIR\_ABUSE\_ALLEGATIONS\_DATA\_TABLE links **abuse\_type\_id** to **sir\_id**.
  + UAC\_SIR\_NEW\_DATA\_TABLE can then be linked to these tables through the **sir\_id** to identify the abuse type of the incident. However, the SIR\_ABUSE\_ALLEGATIONS table only has 18,787 unique SIR\_ID entries (23,385 total entries) while the UAC\_SIR\_NEW\_DATA\_TABLE has 479,070 unique SIR\_ID entries (663,315 total entries). The UAC\_SIR\_NEW\_DATA\_TABLE also has the **abuse\_neglect\_type** column which seems to overlap quite heavily with **abuse\_type** (although not perfectly). 90% of entries, though, have a missing value for **abuse\_neglect\_type**.
* **Programs**: I'll be honest, I am not really sure what these are.
  + LIST\_PROGRAMS\_DATA\_TABLE seems to have the list of programs. These may be agencies or groups that house the unaccompanied children (UAC)? Many of the columns are mostly missing except for the location variables like State and City and the program name. The programs have names like: Ark Homes Foster Family Agency; Board of Child Care - Caminos West Campolina Way; CC Fort Worth Shelter; LSS of New England TIL Group; and Seton Home. This leads me to believe these are where the children are housed or maybe where the significant incident (SI report or SIR) happened?
  + LIST\_PROGRAMS\_DATA\_TABLE can be linked to UAC\_SIR\_EVENT\_DATA\_TABLE through the **program\_id**. THe UAC\_SIR\_EVENT\_DATA\_TABLE is mostly useless though because all the columns are missing. You can use the **date\_created** variable to understand when reports were being filed perhaps. You could also see how many times each program is mentioned in the UAC\_SIR\_EVENT\_DATA\_TABLE to get a sense of how many reports are being filed with each program?
  + PROGRAM\_LEVEL\_EVENTS\_INFO\_DATA\_TABLE seems to contain information about the event or incident, but a lot of the information is redacted. I am also not 100% certain this really has to do with the UAC since the event category types are things like: natural disaster, major disturbance, and death (non-UC) although about 90% of events are classified as other. There are a bunch of binary variables that can then be used to determine specific details about the event: e.g., was there an internal investigation, was law enforcement involved, etc. (although these columns vary in their usefulness). This can be linked to the LIST\_PROGRAMS\_DATA\_TABLE using the **program\_id**.
  + It seems like **EVENTS** and **PROGRAMS** are two key important entities which I am not really sure what they are. It seems like multiple UAC\_SI reports (UAC\_SIR) can be associated with one event. And then events are associated with a program.
* **Medical/Mental Issue**
  + These describe what the medical/mental health issue was that caused the UAC\_SIR to be filed (I think). LIST\_SIR\_MM\_ISSUES\_DATA\_TABLE links MEDICAL\_MENTAL\_ISSUE\_ID to its descriptions or category types (i.e., Childbirth; Death of UAC; Injury Requiring Medical Attention; Medication Issue; or Mental Health Emergency). This table can be linked to SIR\_MEDICAL\_MENTAL\_ISSUES\_DATA\_TABLE which links the MEDICAL\_MENTAL\_ISSUE\_ID to a SIR\_ID. The SIR\_ID can then be linked to the UAC\_SIR\_NEW\_DATA\_TABLE to determine the medical/mental health issue which caused the SIR to be filed. **A very important limitation is that only** 8,433 SIR\_IDs (out of 479,070) actually have a medical/mental health issue associated with them.
* **UAC\_SIR Category**: This basically describes the category of abuse the unaccompanied children experienced.
  + UAC\_SIR\_CATEGORY\_DATA\_TABLE links **category\_id** to its description where possible values include: Other; Runaway; Abuse In DHS Custody; Abuse In Home Country; Abuse In United States (74 possible categories).
  + UAC\_SIR\_CATEGORY\_MAP\_DATA\_TABLE links **category\_id** to **sir\_id**.
  + Using **sir\_id**, one can then link to UAC\_SIR\_NEW\_DATA\_TABLE. Interestingly, the UAC\_SIR\_CATEGORY\_MAP\_DATA\_TABLE has 594,374 unique SIR\_IDs while UAC\_SIR\_NEW\_DATA\_TABLE only has 479,070 SIR\_IDs meaning there are some SIR\_IDs which do not link back the new table.
* **Documents**: Not really sure what this table is. I think it would link the documentation that was officially filed as part of the UAC\_SIR report to its database entry.
  + UAC\_SIR\_DOCUMENTS\_DATA\_TABLE contains information on the type of documents on file (e.g., PNG, PDF) as well as the organization which filed the documents (?) e.g., DOJ, CPS, etc.
  + UAC\_SIR\_DOCUMENTS\_DATA\_TABLE can be linked to the UAC\_SIR\_NEW\_DATA\_TABLE using **uac\_sir\_id**. Not every **UAC\_SIR** entry, though, has documents associated with it, and some document entries do not link to any **uac\_sir\_ids** in the UAC\_SIR\_NEW\_DATA\_TABLE.
* **Notifications**: I think these are the individuals or agencies who were notified when the UAC\_SIR was originally filed?
  + UAC\_SIR\_NOTIFICATION\_LIST\_DATA\_TABLE might have some useful information. The **agency\_title** column seemingly contains the information on who was notified. The most common values include: Case Coordinator; Medical Coordinator; SIR Hotline; ORR/FFS; and CFS. The **notification\_list\_type** seems to contain information on who made the initial report? This is just a guess. Possible values are: ORR; REPORTER CONTACT; OTHER; SAF Others.
  + THE UAC\_SIR\_UCINVOLVED\_LIST\_DATA\_TABLE does not really have a lot of useful information. The only useful column seems to be **uc\_role** which indicates the role of the individual in the case? E.g., they are a victim, perpetrator, reporter, witness, etc.
* **UAC\_SIR\_NEW\_DATA\_TABLE**: The table that links everything together? Unfortunately, so many of the columns are all missing or nearly all missing, it severely limits the utility of this table. Some columns which still contain some information are listed below.
  + An important complication with this data is how exactly all the IDs fit into each other and what rows represent. At this point, I am not sure. Rows are unique on **uac\_sir\_id** and they nest into **sir\_id**, **event\_id**, and **uac\_program\_id**.
  + **IS\_SIR\_FOR\_RUNAWAY**: 77.4% of reports were NOT for a runaway child.
  + **LOCATION\_OF\_INCIDENT**: Possible values are: Other; Housing Area; Medical Facility On-Site; School Area; Recreational Area On-Site.
  + **REPORTED\_TO\_CPS**: 17.7% of cases were reported to CPS.
  + **REPORTED\_TO\_LAW\_ENFORCEMENT**: 2.2% of cases were reported to law enforcement.
  + **SA\_HOW\_UAC\_INVOLVED**: Possible values are: Victim (94.4%); Perpetrator; Witness; or Other.
  + **SA\_TYPE\_OF\_ALLEGATION** and **SA\_TYPE\_OF\_INCIDENT** are about 92.5% missing which is very high, but there are still at least some non-missing entries.
  + **UAC\_OR\_ANYONE\_INJURED**: 12.7% of rows resulted in an injury for someone (?)

Useless tables because nearly all values are missing/redacted and/or the IDs do not link to any other table:

* FOLLOWUP\_CONTACT\_DATA\_TABLE
* LIST\_SIR\_NOTIFiED\_ENTITIES\_DATA\_TABLE
* ORR\_NOTIFICATION\_DATA\_TABLE
* SIR\_OTHER\_INCIDENT\_TYPES\_DATA\_TABLE
* SIR\_UAC\_BEHAVIOR\_DATA\_TABLE
* UAC\_SIR\_FIELDS\_HISTORY\_DATA\_TABLE
* UAC\_SIR\_FOLLOWUP\_RPT\_DATA\_TABLE
  + You'd think this table would be more useful because it has so many columns. However, nearly all of these columns have missing values. A select few do have some information though which will be described below.
  + Another issue with this table is it only has around 1,500 entries meaning it can only be linked to a small fraction of the total number of cases that actually exist. On top of that, I am not really sure what the unit of analysis is in this table i.e., I am not sure what rows correspond to.
  + Based on the table name, this might only be reports that had repeated follow-ups?
  + **LOCATION\_OF\_INCIDENT**: Housing Area, Other, School Area, Recreational Area On-Site.
  + **REPORTED\_TO\_CPS**: 56.3% of cases (?) were reported to the CPS.
  + **REPORED\_TO\_DOJ**: 24.7% of cases (?) were reported to the DOJ.
  + **REPORTED\_TO\_FBI**: 20.2% of cases (?) were reported to the FBI.
  + **REPORTED\_TO\_LAW\_ENFORCEMENT**: 15.9% of cases (?) were reported to law enforcement.
  + **REPORTED\_TO\_STATE\_LICENSING**: 38.7% of cases (?) were reported to state licensing.
  + **SA\_TYPE\_OF\_ALLEGATION**: UC and UC Nonconsensual; Staff and UC; UC Other; UC and UC Consensual; and Non-UC Minor and UC.
  + **SA\_TYPE\_OF\_INCIDENT**: Inappropriate Sexual Behavior; Sexual Abuse; Sexual Harassment; Code of Conduct Violation.
* UAC\_SIR\_MEDIA\_CONTACTS\_TABLE

**April 2024 tables** → These tables exist separately from the above tables as I do not currently see any way in which the April 2024 tables can be linked to the above tables. In general, I think these tables are a bit more useful than the above tables. They seem more straightforward, at least. However, it would be nice to confirm with a domain expert or substantive expert that my hunches about what each of these tables represents is correct.

* **Agency** → Corresponds to the agency involved in investigating the claim of abuse? Unfortunately, the only ID column is **agency\_id** which is not present in any other table.
  + **AgencyName2**: The name of the agency involved in investigating the abuse? Possible values include: Adult Protective Services; Adult Programs Division; Adult Protection Unit; Adult Protective and Community Services Branch. However, it is strange that many agencies are listed many, many times (e.g., Adult Protective Services is listed 67 times). Since Fiscal Year is a category, maybe the unit of observation is agency-year?
  + **Assessment Tool**: 75.8% of agencies use an assessment tool.
  + **Evidence standard**: Possible values include: Preponderance; Credible, reasonable, or probable cause; No state standard; Clear and convincing.
  + **Report Count Accepted** and **Report Count Not Accepted**: How many reports each agency accepted and did not accept? I treated these variables as categorical but perhaps they should be treated as numerical.
  + **Response time**: How many hours does it take for them to respond? Possible values seem semi-continuous thus not truly numerical.
  + Includes a variable for **State** and **Year**.
* **Investigation**:
  + We have **state**, **county**, and **year** columns.
  + None of the other columns have information that appears helpful. I cannot really tell what any of the other columns mean.
  + There is an **investigation\_id** which can be linked to the Perpetrator and Relationship tables, but I do not know how useful this is because this table does not seem very useful. Also, the vast majority of **investigation\_ids** do not appear in Perpetrator or Relationship.
* **Key Indicator**
  + I have no idea what this table is. It has no way to be linked to any of the other tables. Furthermore, there are only 130 rows, and many values for many of the columns do not make much sense. For example, the column **perp\_gender\_female** has values like 0, 110, 130, 170, etc.
  + Perhaps these are aggregated values by year? For example, the number of female perpetrators in a given year. If this is true, I should probably not treat all these columns as categorical variables. I think this explanation makes the most sense.
* **Perpetrator**
  + I think this table has individual-level information on each perpetrator? Like age, behaviors (e.g., alcoholism, anxiety, bipolar, depression, dementia), disability, race/ethnicity, and gender identity.
  + There is also a **state** and **fiscal year** variable.
  + This table contains **Perpetrator Id** which can be linked to the Relationship table. There are some **perpetrator\_ids** which do not appear in the relationship table, though.
* **Relationship**
  + I think this table has information concerning the relationship between the perpetrator and the victim? It’s unclear, though, because not many of the columns contain useful information.
  + There is a **year** and a **state** column.