**Capstone Project Preliminary Analysis Template**

**Name:** Catherine Yarovoi

**Project Title:** *Mental Health in Minors 2020 – SAMHSA Client-Level Data*

**Project Objective:** This project objective is to understand patterns in SAMHSA-reported (and anonymized) minor data for FY2020. Details may include:

* Frequency of co-occurring mental health and substance use disorders in minors (for larger general population, this was part of SAMHSA’s objective in creating this client-level dataset)
* Frequency of co-occurring mental health issues outside of substance use disorders.
* Understanding referral source, details on criminal justice referral, income sources, and residential information where applicable, and any other situational or demographic information for this minor population in 2020

Audience may include mental health practitioners, especially those working with minors, who may want to understand the landscape of minor mental health in the US – this is a snippet (though not exhaustive) of such trends from 2020. Other audiences may include government officials or SAMHSA for purposes of allocating resources to, incentivizing distribution of resources to, or ensuring proper training on how to service areas and communities where assistance is most needed and used.

**Data Collection Source:**

* Codebook includes full methodology description: [*https://www.datafiles.samhsa.gov/sites/default/files/field-uploads-protected/studies/MH-CLD-2020/MH-CLD-2020-datasets/MH-CLD-2020-DS0001/MH-CLD-2020-DS0001-info/MH-CLD-2020-DS0001-info-codebook.pdf*](https://www.datafiles.samhsa.gov/sites/default/files/field-uploads-protected/studies/MH-CLD-2020/MH-CLD-2020-datasets/MH-CLD-2020-DS0001/MH-CLD-2020-DS0001-info/MH-CLD-2020-DS0001-info-codebook.pdf)
* The data from MH-CLD and MH-TEDS are for individuals receiving mental health treatment services provided or funded through state mental health agencies (SMHAs). SMHAs are the state entities with primary responsibility for reporting the data.
* MH-TEDS is focused on treatment events, such as admissions and discharges from service centers. Admission and discharge records can be linked to track treatment episodes and the treatment services received by individuals. Thus, with MH-TEDS, both the individual client and the treatment episode can serve as a unit of analysis. In contrast, with MH-CLD, the client is the sole unit of analysis.
* MH-TEDS enhances the ability to report data on people with co-occurring mental health and substance use disorders. MH-TEDS also offers optional data fields for individuals with mental illness that are not captured in MH-CLD, such as referral source, details on criminal justice referral, income sources, and health insurance.
* Data does not include all individuals receiving mental health treatment services.

**Methodology:**

* I first looked at the dataset in Python to get an idea of its original structure. I did the standard data checks (shape, info, dtype, null values). Note that null values are replaced with a “missing or not provided” field in this dataset.
* After seeing that I could work with the data, I used Python to rename the columns from acronyms to more descriptive column names in order to make the data easier to work with. I then tried to do some data manipulation, but Jupyter kept crashing because of the dataset size (almost 7Million rows). I couldn’t run Python long enough even to split the data into smaller datasets, so I took it into BigQuery and Google Cloud to query the results.
* The data is organized on a client-level basis and almost every column contains several variables created from the original variables submitted by the states. For example, a variable was created to indicate whether a given mental health diagnosis was reported in the first, second, or third diagnosis field.
  + Because of the large file size, I first filtered the data to under18 years of age using the “AGE” column and referencing the AGE variable information provided in the data dictionary and codebook for this dataset to understand how to filter the data.
  + I then used many “CASE WHEN” statements to clean the data, referencing the variable descriptions provided by the dataset’s codebook.
  + One of the limitations I ran into was that BigQuery did not allow me to change the datatype without creating a new column (for a new datatype) unless I wanted to recreate the dataset with a new, manually set schema. I therefore could not use a “set” function instead of creating new columns with “case when”, as I could not set an integer datapoint (the variable number) as a varchar datapoint (the updated mapped description of the variable).
  + After cleaning the dataset, I then had to delete the original columns with the variables only and keep the new columns I had created with the mapped information.

**Findings:**

1. **Data Size:** *How many rows and columns?*1,915,199 rows by 40 columns
2. **Missing Values:** *How many? Will these need to be replaced or deleted?* N/A
3. **Duplicates:**  *Describe how this was detected None;*

SELECT

    \*, COUNT(\*)

FROM

    `mh\_analysis\_updated.mh\_under18\_labeled\_fulldata`

GROUP BY

    YEAR,NUMBER\_OF\_DIAGNOSES,TRAUMA\_STRESS\_DISORDER, ANXIETY\_DIS, ADHD\_DIS, CONDUCT\_DIS,DIMENTIA\_DIL\_DIS, BIPOLAR\_DIS, DEPRESSIVE\_DIS, OPPOSITIONAL\_DEFIANT\_DIS, PDEVELOPMENTAL\_DIS, PERSONALITY\_DIS, SCHIZOPHRENIA\_or\_PSYCHOTIC\_DIS, ALC\_SUB\_DIS, OTHER\_DIS, CASEID, age\_range, educ\_category, ethnicity\_category, race\_label, sex, state\_psych\_hosp, SMHAfundop\_COMM\_PROGRAM, served\_other\_psych\_inp, res\_treat, under\_justice\_system, mh\_diag\_1, mh\_diag\_2, mh\_diag\_3, SUD, marital\_status, SMI\_SED, SubstanceU\_Prob, employment\_status, not\_in\_labor\_force\_detailed, vet\_status, residential\_at\_discharge\_end, reporting\_state, census\_division, census\_region

HAVING

    COUNT(\*) > 1;

1. **Dates:***What is the date range? Will this need to be modified for the project?* **Fiscal year 2020** (end year of the state-defined reporting period, such state-defined reporting periods, are mostly July 1, 2019–June 30, 2020, but in some cases are October 1, 2019 –September 30, 2020 fiscal year, a lagged fiscal year, or both), **not modifying.**
2. **Additional Findings:** *Identify any other findings or errors in the data. Any outliers? Any inconsistencies relative to the project? There is a fairly large amount of “missing or not provided” pieces of information, especially among the minor population of this dataset. I may not be able to dig as deeply into some of the demographic information as I would like for minors specifically because of this.*

**Attach any SQL or Python code in a separate file**

Link to BQ query:

* <https://console.cloud.google.com/bigquery?sq=805692182138:a5b7ebc5660846f08c1d739e49fa8770>
* [*https://console.cloud.google.com/bigquery?sq=805692182138:f2eb038584454ebb8e27ef394d5c7e20*](https://console.cloud.google.com/bigquery?sq=805692182138:f2eb038584454ebb8e27ef394d5c7e20)

*DBeaver SQL also attached if having issues with links above*