**Capstone Project Proposal**

**Name:** Catherine Yarovoi

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

**The Data**

The Mental Health Client-Level Data (MH-CLD) and the Mental Health Treatment Episode Data Set (MH-TEDS) systems provide information on mental health diagnoses and the mental health treatment services, outcomes, and demographic and substance use characteristics of individuals in mental health treatment facilities that report to individual state administrative data systems.

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.

**Project Objective**

Using the MH-CLD client level data, this project aims to understand the use of treatment services, demographics, and characteristics of minors (children under 18 years of age) serviced through state mental health agencies. The project also assessed the co-occurrence of mental health diagnoses for this population.

This project is aimed at providing state agencies with more information on the populations they are serving by understanding mental health needs and thus assisting in state representatives to be able to properly allocate or assess state-funded services.

**Previous Research or Current Knowledge**

This is my first time working with client data. I have some exposure to research around the topic of mental health needs, having worked on several national communications campaigns around suicide prevention for youth (see [SeizeTheAwkward.org](https://seizetheawkward.org/)).

**Project Design and Workflow**

1. Python to first assess the data and understand basic data structure (the file is almost 7M rows and will not open in Excel). Standard data checks were performed in Python (Jupyter Notebook), such asshape, info, dtype, null values, duplicates. I noted that all missing values are not reported as null but rather as a “missing or not provided” string value in this dataset.
2. Data columns were renamed for ease of reference in Python. I later realized this step could be avoided as the columns were later renamed through a WHEN CASE statement in SQL, so if repeating this process I would not initially rename the columns in Python. However, this step did get me more familiar with the data.
3. Because of the size of the file, I brought the file into Big Query and stored the initial data in Google Cloud rather than working in Python on my local machine.
4. I decided to focus on children/minors to decrease the size of the dataset. I used SQL and referenced the codebook to filter the data.
5. 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 which mental health diagnosis the client received as their primary mental health diagnosis. For example, the variable one in column MH3 indicates that the primary mental health diagnosis for client X was a Trauma or Stressor-Related disorder. In order to make this data interpretable and readable, I created a SQL query using many “CASE WHEN” statements, referencing the variable descriptions provided by the dataset’s codebook.
   1. Note that if repeating this process, this is where you can relabel the dataset columns as well.
6. One of the limitations I ran into during the cleaning process was that BigQuery does not allow changes to the datatype without creating a new column (for the new datatype) unless you recreate the dataset with a new, manually set schema. Therefore, because the variables from the dataset were initially registered as integers and I wanted to transform them to string variables, I could not use a “set” function instead of “case when” statements. Using “set” would have allowed me to replace the integer column with the string column, but “set” does not allow setting an integer datapoint (the variable number) as a varchar datapoint (the updated mapped description of the variable).
7. After cleaning the dataset, I then had to delete the original columns with the integer variables and keep only the new columns I had created with the mapped variable information.
8. This file was brought back into Python for a few data quality checks and an easier export process.
9. The file was then exported and brought into Tableau. I did much of my analysis in Tableau, cross-checking the data in Python.
10. Separately, I also wanted to understand diagnosis co-occurrence. This required me to run either a chi-square analysis or do a product matrix to asses the different diagnosis and how often they co-occurred in the dataset. The various diagnosis are all binary categorical variables.
11. The easiest solution that I found was to isolate these 13 variables from the dataset (representing 13 diagnoses) and create a product matrix in Python, which I could then analyze in Excel.
12. I completed the analysis of the product matrix in Excel, and saved my two heatmap tables as images to be able to bring them into Tableau for a cohesive report.

**Data Collection**

Data was downloaded as a delimited file from <https://www.samhsa.gov/data/data-we-collect/mh-cld-mental-health-client-level-data>

The data comes with a codebook, which was used to decipher the variables in the data: [*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)

One section of the analysis also looks at population information, and thus includes data from <https://cwoutcomes.acf.hhs.gov/cwodatasite/population/index>

**Intended Outcome**

The intended outcome for this project is to identify demographics and characteristics of minors (children under 18 years of age) serviced through state mental health agencies, as well as service usage by state. An interactive dashboard allowing state medical providers or state representatives to filter to their state is part of the final design.

On a country-level, the project also aims to identify trends in mental health diagnoses, such as co-occurrence of diagnoses.