My project will be to combine the datasets of a website, an API, and a csv file which are all about schizophrenia. I plan to then take my consolidated dataset and use it to make a screening tool to determine which medications may be best for people of certain demographics.

**Flat file:** Psychiatric\_Drug\_Reviews\_WebMD.csv

This file is 13 columns wide by 61,320 rows long. The file has information on patients and caregivers, type of mental illness, age group, symptoms, and medications.

**Website w/ table:**

<https://ourworldindata.org/grapher/schizophrenia-prevalence-by-age?tab=table>

This table is 16 columns wide by 222 rows long. It contains information about the prevalence of schizophrenia in specified age groups in various countries.

**API base Sandbox url:**

<https://sandbox.api.service.nhs.uk/nhs-website-content/>

API Mental health feelings symptoms and behaviors:

<https://sandbox.api.service.nhs.uk/nhs-website-content/mental-health/feelings-symptoms-behaviours/>

API Mental Health Conditions:

<https://sandbox.api.service.nhs.uk/nhs-website-content/mental-health/conditions/>

API Mental Health:

<https://sandbox.api.service.nhs.uk/nhs-website-content/mental-health/>

API Medicines:

<https://sandbox.api.service.nhs.uk/nhs-website-content/medicines/>

All these links are from the same API, just different sections of it. It has information about symptoms, medicines, and potential causes of schizophrenia.

The website table relates to the csv file as they both contain information about age groups of those with schizophrenia. The API also relates to the csv file because it has information about medications and symptoms like the csv file also has. So while there is not a direct relationship between the API and website they have a common link in the CSV file.

For this project, I intend to group data by information each set has in common with the CSV file. The CSV file relates to the other two datasets based on the age range for the website's table and by medications for the API. I think there will be a bit of a struggle trying to match up extracted data. I know the data relate to each other through the CSV file, but I'm unsure how it will look once I've pulled the relevant data from each source. I'll be subsetting the data from the CSV and table; I'm still determining what the information from the API is structured like. Because I do not know the structure of the data from the API, I'm unsure what issues that may cause. Once I have the relevant data matched up, I can then append the dataframes together. Once all the data is linked together, I can start to identify trends and potentially build a model.

There could be ethical implications for my end goal with these datasets. Given their attributes, I want to use this data to build a model that could predict which medication might be the most successful for someone. For example, a female, age 45-50, with some specific symptoms may have a different recommended medication than a male in his 20s with different symptoms. To truly make a model like that, I would need significantly more data about the people I'm training the model from. I have a few other datasets that might be useful to incorporate down the line, but there are some ethical implications of collecting that much data about anyone. Still, ethics also comes into play when determining if the model is good enough to make a prediction rather than a doctor. What if it makes the wrong prediction? Who is responsible for the outcome of the patient? If a doctor chooses to use this model, are they ultimately responsible for the consequences, or is it the model developer?