Title: Impact of Mental Health on Marriage in the USA (2021 and 2022)

Main Question: Does mental health, specifically mental illness and suicidality, impact marriage rates in different USA states?

1. Question Background:

In recent years, mental health has become an important topic in the United States. Many people are struggling with mental illnesses or having thoughts of suicide. At the same time, marriage rates in some states seem to be going down. Could these two things be connected? This project will explore whether mental health problems, like depression or serious thoughts of suicide, are related to fewer people getting married in the U.S in each state.

1.1. Scope of the Question

The project focuses on specific aspects of mental health and suicidality. Instead of addressing all mental health conditions, we narrow the focus to these key factors:

Category	Factors
Mental Health	1- Any Mental Illness 2- Serious Mental Illness 3- Received Mental Health Treatment 4- Major Depressive Episode
Suicidality Factors	1- Had Serious Thoughts of Suicide 2- Made Any Suicide Plans 3- Attempted Suicide

limiting the scope to these measurable and significant indicators, the project aims to provide a focused analysis on how these factors relate to marriage rates in the United States based on each state.

2. Data Sources

The datasets come from two reliable sources: the Substance Abuse and Mental Health Services Administration (SAMHSA) and the National Center for Health Statistics (NCHS) in collaboration with the Centers for Disease Control and Prevention (CDC). Both organizations have their own terms of use, and an email was sent to each for clarification. Details about their policies are provided in the tables below.

Dataset 1: Marriage Rates by State (2019-2022)

Dataset	Details
Metadata URL	CDC Marriage Rates by State
Data URL	Downloadable PDF
Data Type	PDF
License	The data from (NCHS) and CDC is protected by strict privacy laws. These laws allow the data to be used only for statistical reporting and analysis. As long as these rules are followed, it is permissible to use the data the rules are: 1. **Use for Analysis Only**: The data can only be used for statistical purposes like identifying trends or patterns. 2. **Privacy is Protected**: All personal identifiers are removed to ensure confidentiality. The full policy is available here, and **the project complies** with these conditions by focusing on statistical analysis, not identification of participants.
Description	This dataset shows the number of marriages per 1,000 people in each U.S. state for 2021 and 2022. It allows for state-by-state comparisons of marriage trends to identify patterns and explore their potential correlation with mental health factors.

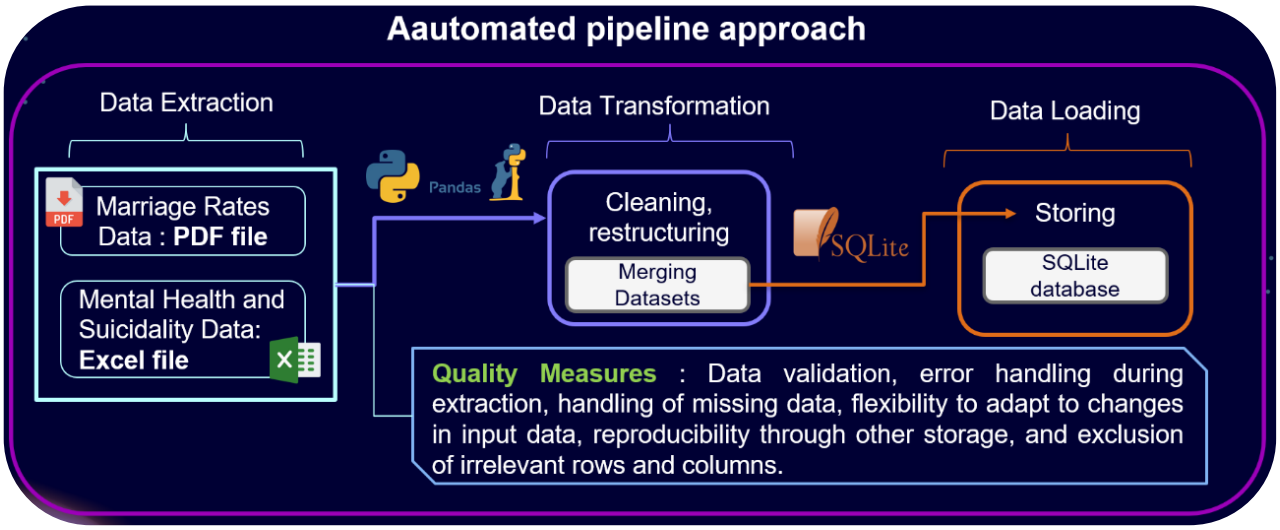
Dataset 2: Mental Illness and Suicidality (2021-2022)

Dataset	Details
Metadata URL	SAMHSA Data (Substance Abuse and Mental Health Services Administration)
Data URL	Mental Health Data (XLSX)
Data Type	Excel Tables (XLSX) / 7 sheets (7 tables)
License	The data I used falls under the **Freedom of Information Act (FOIA)**, which allows access to federal agency records to promote transparency and public understanding of government operations. The data is publicly accessible confirming its availability for use under FOIA guidelines. Additionally, I sent an email inquiry to SAMHSA, and they clarified that while there are exceptions to FOIA that restrict access to certain privileged or confidential information, the data I accessed does not fall under any of these exceptions. Thus, it is openly available for research and complies with all applicable regulations. *And here is the link they provided click here*
Description	This dataset includes state-level data on mental health and suicidality across different age groups and metrics. For each indicator, the dataset provides: - **Age Groups**: Includes specific data for **12-17**, **18-25**, **26+**, and **18+ (all adults)**. - **Estimates**: The percentage of individuals in each age group experiencing a particular mental health condition or suicidality factor. - **Confidence Intervals**: Includes upper and lower limits for each estimate, representing the 95% confidence range. - **Indicators**: Covers metrics like Any Mental Illness, Serious Mental Illness, and Suicidality Factors (e.g., Thoughts of Suicide, Suicide Plans). **Note:** The data is structured but requires cleaning to remove unnecessary columns (e.g., confidence intervals) and rows that do not add value to the analysis. This will be detailed in the next section.

3. Data Pipeline

pipeline on a high level

The pipeline for this project is implemented in **Python**, using libraries such as **pdfplumber** and **requests** for dataset extraction, **pandas** for data transformation, and **sqlite3** for database integration. Error handling is implemented using built-in Python capabilities **(try-except blocks)** to manage issues during data extraction, transformation, and database operations. The figure below shows the pipeline.



Transformation or cleaning steps

Purpose: Clarity in reading such that data used in computations is consistent and avoiding mistakes.

For extraction: requests and pdfplumber library are used for fetching marriage rates from a static PDF. While Read Excel sheets using pandas, skipping irrelevant tables and focusing on critical data points for mental health indicators.

For Transformation: The columns were renamed as there were **7 tables** and the header names were the same so it was a must to rename it for clarity and consistency. Adding to that there were young age groups and extra statistics measures columns which are irrelevant columns, so I dropped it, such as confidence intervals as it is totally logic if the interval is small then there is no need to take it into the account. Also Handled missing values by removing rows with entirely null data. Last but not least make sure the data has the correct assign to **correct data type**. Finally Converted percentages as after extraction they changed as they gave the result divided by 100 so I had to make sure they will appear the same in loading for better readability.

Finally for Loading: Used sqlite3 to initialize and store in an SQLite database named and loaded the data into a structured table, Mental_Marriage_Data, for easy querying and analysis.

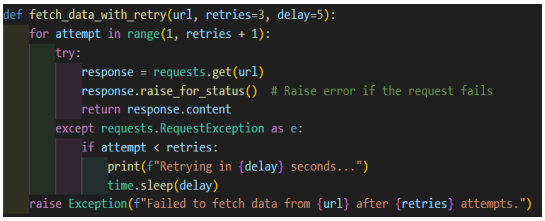
- Problems encounter and solutions

Challenge	Solution
Dataset fetching	Static link extraction for PDF data. Avoided dynamic content issues. As the excel link was dynamic and the pdf link was static for Marriage Rates by State which was challenging to parse it via scrap libraries like selenium and beautiful soup
Data change	Some data did change after extraction resulting in giving result of division, so I had to return it as it was before by reversing the operation
Managing errors	Added error handling (try-except blocks) and detailed logging for invalid inputs.

- Meta-quality measures and deal with errors or changing input data

The pipeline addresses inconsistencies in data formats and incorporates meta-quality measures to ensure reliability. For example, extracting data from a static PDF required custom patterns for parsing, while Excel sheets with varying row structures were standardized by dropping irrelevant metrics like confidence intervals and younger age groups. Missing data was managed by removing rows with critical nulls, ensuring consistent computations while raising errors and stop the pipeline if whole sources were not reach after **3 times** of retry to fetch the data for example.

Hence meta-quality measures included data validation checks, logging of transformation steps, and dynamic column alignment to adapt to future changes in data structure. Additionally, database reinitialization prevented conflicts between outdated and updated datasets, maintaining data integrity during storage.



4. Result and limitations

The result of the pipeline is a SQLite database (merged_mental_marriage_data.sqlite), containing a table named (Mental_Marriage_Data). The table combines marriage rates (from a PDF) and mental health indicators (from an Excel file) for U.S. states in 2021 and 2022. The data is clean, consistent, and easy to use, with proper formatting, renamed columns, and missing values handled.
Hence the reason SQLite database format was chosen because it is lightweight, open-source and easy to query.

- **Critical reflection towards Limitations:**
  - One US state was missing: The data was not complete from the Substance Abuse and Mental Health Services Administration.
  - Age Group Overlap: The data provides percentages for three broad age groups (18+, 18–25, and 26+). These categories may overlap when analyzing trends across groups, complicating detailed insights.
  - Cross-Year Comparability: The dataset combines data from 2021 and 2022, but it doesn't explicitly address how changes between these years could impact trends or correlations.