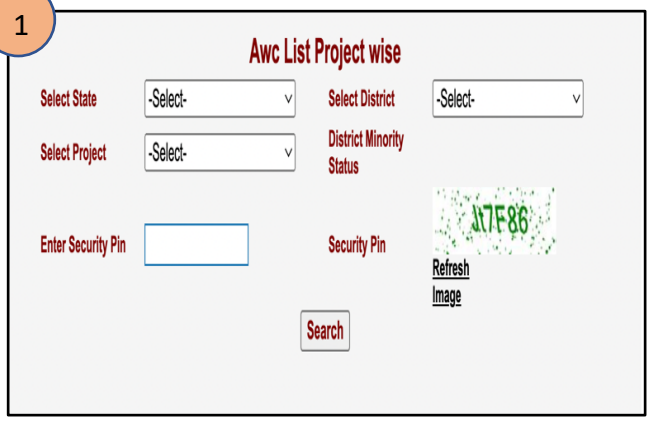
Data from Management Information System of Nutrition Centers using Python

Skills: Web Scraping and Automation

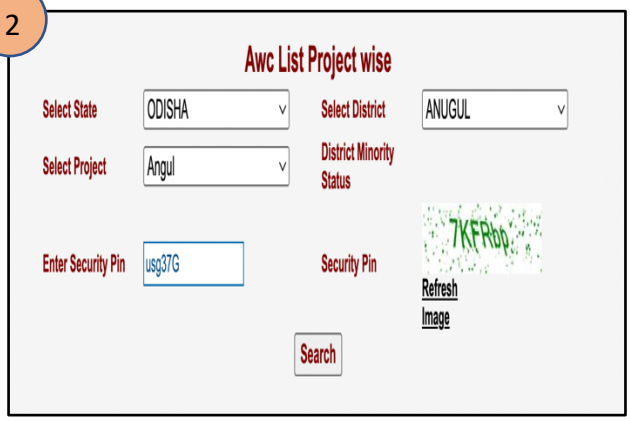
Web scraping is a valuable tool in social development as it allows for **efficient data gathering from websites and online sources**, which can inform decision-making and program evaluation. However, accessing local center-level data in India can be challenging due to data being stored in **non-standard formats or spread across multiple website**s. Therefore, web scraping using Python can help to automate data gathering, saving time and resources. Python libraries, such as **Beautiful Soup**, can **extract data from HTML and XML documents**, navigate website structures, and store data in a structured format.

Task	Details	Challenges	Solution
Scraping and extracting data in a structured format	Deployed the scraping solution using the selenium web driver in Python to extract the data from nutrition centers in a district from a public government MIS website	The captcha security applied at the website for each center restricts the code to automatically fill the form and open the link to the data. I had to enter the captcha manually for each center to extract the data.	Extract center codes at the project level, and then open the link for each center by appending the state, district, project, and center code iteratively to extract the data for all the centers in one project.

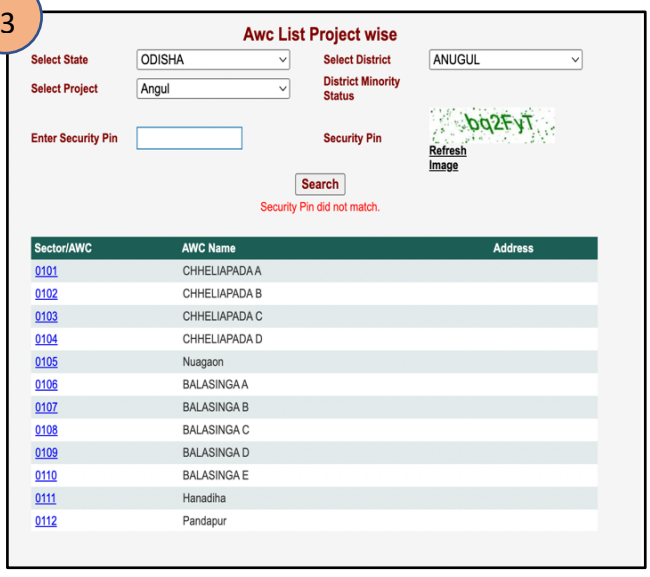
1



2



3



4

Impact Evaluation and Causal Analysis using STATA

Skills: Data Cleaning, Transformation, and Causal Analysis using Probit Regression

The impact evaluation aims to establish a 'proof of concept' for a digital intervention that **generates demand for Sexual and Reproductive Health (SRH) services**. For each survey round, a **multi-stage PPS systematic sampling method** was used to estimate a **sample of 1700 young women** between the ages of 16 and 24. The focus of interest for this study is the **level of knowledge among young women regarding key SRH topics** and issues. Additionally, the study examines the proportion of young women who intend to access SRH services, as well as their access to an SRH service within the last six months if needed.

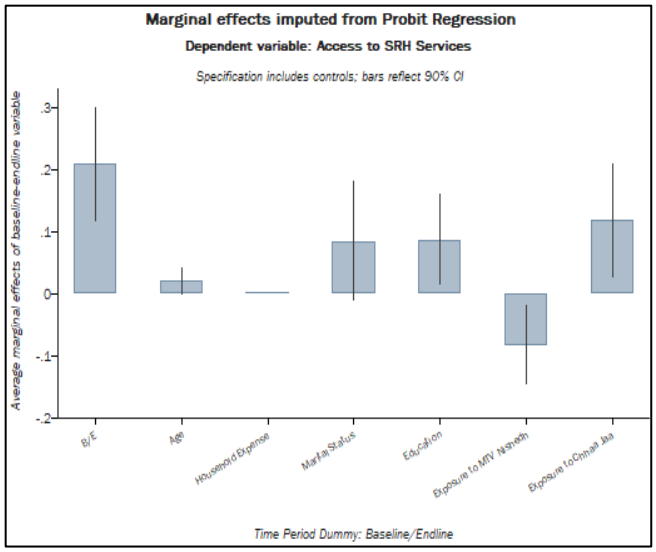
***Probit specification for binary outcome variables*:**

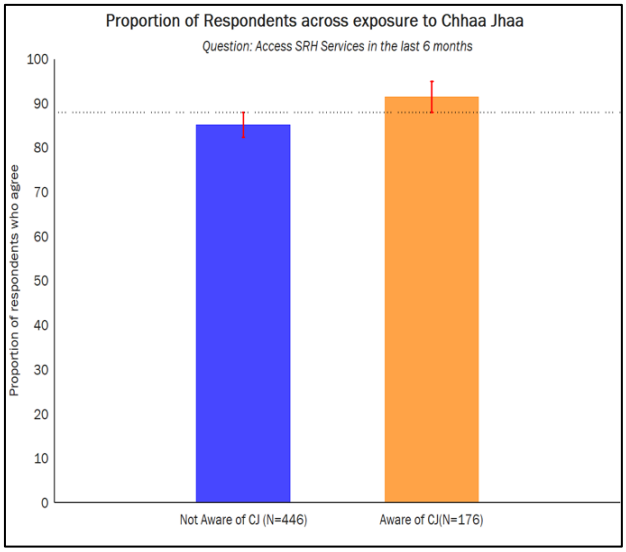
$$P(Y_i)=\alpha+ \beta_1*\textit{Pre-Post Dummy}+\beta_2*\textit{Demographic Controls}+ \beta_3*\textit{Knowledge Controls}$$

To evaluate the impact of the program, a **probit regression specification model** is employed, which includes a pre-post dummy variable. The average marginal effects of the variables are estimated to produce effect sizes that are straightforward to interpret. Impact assessment tests are conducted using a 10% critical p-value while incorporating controls to account for any confounding factors.

The data set is subjected to several data cleaning processes to prepare it for regression analysis. This includes creating smaller subsets based on research criteria, dropping irrelevant or duplicate variables, recoding numeric variables according to specified rules, re-categorizing categorical variables into broader categories, dropping non-significant observations, creating new composite variables, renaming or labeling variables, and conducting sample balance tests. By performing these steps, the data set is structured and analyzed appropriately to generate meaningful results.

Results: A significantly higher proportion of respondents accessed services for SRH issues at the end line.





Heatmap that shows the incidence of malnutrition at the district level in India

Skills: Data Visualization (using ggplot), Merging and Working with Shapefiles.

India currently faces challenges with data-driven decision making due to data existing in silos with **no unique common identifier for merging data sets**. The lack of common identifiers results in a loss of cross-sectoral insights, and lessons and practices are not uniformly applied. To overcome this issue, it is necessary **to map datasets from their native identifiers into a set of common identifiers**, enabling **merging across time, space, and sector**.

The district-wise heatmap below provides a use case that identifies areas with high incidence of malnutrition i.e. **Low Birth Weight** (National Family Health Survey 2015-16), revealing that the patriarchal society of the Hindi belt in the north has poorer outcomes, while the north-east and south have relatively better outcomes. India's shapefiles have 766 districts mapped to the Local Government Directory, while the National Family Health Survey is mapped to Census, which is an older system and only has 640 districts.

The discrepancy in mapping between the Local Government Directory and Census highlights the need to **map all public domain data to a common identifier** to enable decision-makers to draw accurate cross-sectoral policy insights to tackle malnutrition effectively.