**Data Source Overview:**

For this project, I utilized a dataset from the Nepal Demographic and Health Survey (DHS) conducted in 2022. The DHS Program is managed by the United States Agency for International Development (USAID), and it carries out household sample surveys in low and middle-income countries to monitor changes in population, health, disease patterns, and nutrition over time. The latest DHS survey in Nepal was conducted in 2022 and aimed to gather comprehensive data on various aspects of population demographics, health indicators, and healthcare access in Nepal.

The dataset can be found on the official DHS Program website (dhsprogram.com) but requires a special request and a proper proposal with a defined team member to access it. Although the program provides open access to survey data for researchers, policymakers, and organizations, it is not allowed to be used for the undefined project and shared among others. Access to the dataset enables analysis, policy formulation, and program evaluation related to public health and development in Nepal.

**Fit for Use Criteria and Evaluation of Data Fit for Use:**

I used the research question to determine the required data variables. This study aimed to explore the distribution of non-communicable disease (NCD) risk factors, such as blood pressure and body mass index (BMI), among female DHS respondents and their association with sociodemographic variables in Nepal. This helped me define the variables for my project, and I selected 42 variables after a literature review for this study. I defined the criteria for the fit of my data as follows:

Completeness: We consider the data in our dataset to be complete if it contains data for our outcome of interest, i.e., BMI and hypertension. If any data is missing in these variables, we exclude it. Furthermore, we defied our population of analysis to be more than 18 years of age so even if the data was complete the under-18 of age group data was excluded. This is our main fit-for-use criterion, which ensures no missing values in critical variables, and the dataset covers all necessary aspects of the research question.

Consistency: Variables must follow a consistent format across the dataset and have coherent values across all the variables. For example, the dataset for blood pressure must not present decimal point values and must follow a similar pattern across all the data sets. In contrast, the flexibility was given to BMI as it can be reported as an integer. Missing values without the way of leveraging imputation will be removed, and it must fall under the standard provided in the guidebook.

Relevance: For our research question, we defined 42 variables included in the dataset that are relevant to the research objectives. These variables were relevant according to previous literature. As the dataset contains more than 6000+ variables across, we only extracted the 42 predefined variables as they were more relevant to our research question. Several data from the dataset were excluded, such as the men questionnaire, kids questionnaire, maternal and child health questionnaire, mental health questionnaire, accident questionnaire, and so on.

Fitness: The final dataset will meet specific criteria, only if there are no missing values across any variables it will be considered fit for use in our research. Our final dataset will contain only the rows that have all the values across all the variables to be considered as a fit or standard set for the project. As it is survey data, we do not expect the data to be distributed normally across variables. Methods such as visual plots of histograms, as well as box and whiskers plots, were performed to assess the fitness for further use of the data and the determination of the statistical method.

**Cleaning Steps:**

**On R:**

I have processed a dataset by merging two different datasets, filtering the data, and imputing missing values. I used R for merging and variable selection and Excel for data imputation. The R libraries that I used include tidyverse, knitr, mlr3, readr, dplyr, haven, gtsummary, foreign, survey, labelled, and forcats.

To clean the data, I used the read\_dta function to read the NPPR82FL.DTA file containing the whole survey dataset. Similarly, I used the read\_dta function to read NPIR82FL.DTA to read the women's dataset from the survey. I extracted data labels from both datasets to understand variable names and used the standard data recode manual to understand the data better.

For the women's file, I selected variables such as case identification, cluster number, and household number, along with demographic information such as age, marital status, religion, ethnicity, and education level. I also included variables related to health indicators such as weight, height, body mass index (BMI), smoking behavior, and blood pressure. These variables are essential for examining the distribution of non-communicable disease (NCD) risk factors and their association with sociodemographic characteristics among female DHS respondents in Nepal.

Similarly, from the all files, I selected identifiers and demographic information for households and household members. Key variables related to health indicators such as weight, height, BMI, smoking behavior, and blood pressure measurements were also included, along with additional variables such as wealth index and education level. These variables provide a comprehensive dataset for analyzing the prevalence and distribution of NCD risk factors among female respondents and their households in Nepal.

To merge the data files, I renamed the identifier column names to match the name across both files. I used the DHS recode manual and website as a guide to rename the columns in the household data to match those in the women's data. To ensure compatibility for merging, I converted the identifiers in both datasets to the same data type, which was numeric in our case. Then, I sorted both datasets based on the common identifiers to facilitate the merging process.

The datasets were then merged using the `merge()` function, specifying the common identifiers (`v001`, `v002`, `v003`) for the merge operation. The parameter `all.y = TRUE` ensures that all rows from the women's data are retained in the final merged dataset, even if there are no corresponding matches in the household data. Overall, this segment prepared and merged the household and women's data, aligning them based on common identifiers and ensuring the compatibility needed for subsequent analyses. Finally, I extracted the file in CSV format and processed it in Excel.

**Further Process In Excel:**

To streamline the data and prevent redundancy, I ensured that only one set of each duplicated variable was preserved in the final dataset. Then, I implemented filtering and cleaning procedures to improve data quality. Specifically, I removed 1,661 rows where respondents were below 18 years old and eliminated 9,670 rows containing missing data in blood pressure-related variables. To address missing values in critical variables, such as "years since first cohabitation" and "smoking status," I used data imputation techniques. For example, I set the "years since first cohabitation" variable to 0 if respondents were unmarried, and replaced missing values in smoking status with "NO" as there were only values present for "Yes." to bolster the completeness and usability of the dataset, minimizing the impact of missing information on our analyses.

The resulting dataset, saved as "merged\_womens.csv," comprises 3,516 rows and 42 variables, encompassing the fundamental aspects pertinent to our research objectives. After addressing any remaining missing values through imputation techniques, the dataset was further refined to 3,513 rows while retaining the same 42 variables. This final dataset is ready for comprehensive analysis, facilitating robust insights into the distribution and associations of non-communicable disease risk factors among female respondents in Nepal.

**Documentation of Cleaning Steps:**

To document the cleaning steps taken, I have created a README file that includes the description of the dataset, the variables it contains, as well as R code, I have saved files in standardized formatting such as .csv format for the dataset and .txt file for readme file. Similarly, the variable names are provided in the .csv data label file names.

**Data Analysis:**

At present, due to time constraints and being the only member of the project, I have only analyzed the normality distribution of the main variables, i.e., Blood pressure, BMI, and age. However, later on, I will analyze the dataset to answer the research questions. I will use statistical techniques and visualization tools to explore the distribution of NCD risk factors among female DHS respondents in Nepal. The analysis will assess whether the cleaned dataset provides sufficient insights and identify any further steps needed to address the remaining research questions.

**Result of Initial Analysis:**

**A collage of graphs

Description automatically generated**