**HRS Data Analysis - Vidyun**

The Previous work on HRS data is conducted and it is used as a reference for data gathering process. There are some changes made in data where some variables are replaced and new variables are added. Below are the files and process it follows to reach the quality data and visualization we reached.

**Data:**

The Data needed for analysis is available in this [link](https://hrsdata.isr.umich.edu/data-products/rand?_gl=1*1uuvviq*_ga*MTUyNDA3MTE3NC4xNjk2NTMzODMx*_ga_FF28MW3MW2*MTcwMTM2MjU0My40MS4wLjE3MDEzNjI1NDMuMC4wLjA.). The data from 2006 to 2016 are downloaded and used. It is better to use a .sav file as other formats store the same data in a bigger size. This can reduce space or memory issues while loading data.

**Generating Dataset:**

The Re\_Data\_generate\_10\_19.ipynb file houses essential code for generating a consolidated dataset in CSV format, suitable for subsequent analysis. Despite each yearly dataset boasting over 1500 columns, we selectively extract only the pertinent ones for analysis, as outlined in the accompanying data dictionary available [here](https://hrs.isr.umich.edu/documentation/question-concordance).

Given that each year's data initiates with a distinct prefix, adhering to a constant format, it's noteworthy that the questions posed may evolve, leading to alterations in data variable names. To accommodate these variations, we systematically rename the columns for each year before amalgamating them to craft the final CSV file. A meticulous cross-verification process is conducted to ensure accuracy, involving a check on the total number of records at the conclusion of the merging procedure.

**Data Cross checking and Verification of Availability:**

In the file Data cross check after meet.ipynb, we check if the data is collected successfully by checking the descriptive statistics of each year's data with the documentation on the website.

This helped us find the variable data missing for each year. In addition, We also check the [Questionnaire](https://hrs.isr.umich.edu/documentation/questionnaires) and how the question is framed, Decoding the values.

**Analysis and Data Cleaning:**

The data cleaning process is documented in the files Composition Analysis.ipynb and cleaned data arrange.ipynb. One of the primary challenges encountered in the project was dealing with missing data, particularly in the context of the Leave-Behind questionnaire. We define data as missing when all information pertaining to a specific category (e.g., Discrimination, Trauma) is absent.

Ensuring uniformity in the direction of Ordinal data (More - Worse) is imperative. Additionally, for certain variables, we identify composition score as missing if it exceeds a specified threshold of missing variables as indicated in the documentation.

Rather than computing the composition score by averaging across all questions provided, we adopt an approach that considers the number of questions answered. This nuanced method provides a more accurate representation.

In instances where data is absent for certain years, a normalization process is implemented to bring values into a standardized range. For example, if Discrimination Q6 is missing for 2006 but available in other years, we opt to calculate the average based on the available five questions and subsequently standardize the values.  
  
In the end, the file is exported into csv format to reduce re-running of code.

**Visualization:**

The Univariate and Bivariate.ipynb file is dedicated to visualizing relationships between different variables. Our analysis approach varies for categorical and continuous variables.

For categorical variables, we delve into the number of missing values, mode, count of unique values, and their distribution. Meanwhile, for continuous variables, we extract descriptive statistics to gain a comprehensive understanding.

To unravel relationships, we calculate the average of a continuous variable (X) for each value of another continuous variable (Y), considering Y as an ordinal variable. This approach unveils trends within the data. When comparing continuous versus categorical variables, we employ boxplots to discern patterns effectively.

Noteworthy visualizations include:

1. Examination of cognitive function - objective versus subjective (self-memory versus composition score for cognition).

2. Exploration of the influence of race and ethnicity on cognitive impairment.

3. Analysis of cognitive impairment against all other composition scores.

These visualizations serve as crucial tools in unraveling intricate relationships and patterns within the dataset.