ETL REPORT

Background

For project 2, our group decided we wanted to use data collected by a Stanford University research study to create an interactive portal where people can explore how couples meet and stay together. With that lens in mind, we planned how to extract, transform and load the data into a database that would lay the groundwork for the project.

The Data

Stanford's How Couples Meet and Stay Together research data is available publicly for download on the project website in the form of stata (.dta) files. There is an initial data file containing the baseline survey results, and two of the subsequent 5 follow up survey results (baseline/wave 1, wave2 and wave 3). There are additional stata files for each of the other follow up surveys (wave 4, wave 5, and wave 6). Each wave consists of demographic survey questions, and then questions around the individual and their relationship

Process

Step1: Understanding the research study and the data
We took a fair amount of time to work backwards from the end goal of what we wanted the portal to look like, and mapping that back to the structure of the research study and the data. One issue we needed to resolve was around when academic rigor was not necessary for our purposes. For example, demographics such as gender, age and race were asked every time the survey was administered even though they are not likely to change. While this may be necessary for academics, it is unnecessary for our purposes. We also needed a good understanding of columns and how they related to each other, especially when they occurred in many waves.

Step 2: Clean/Transform

- Load stata data using pandas function and convert to dataframe and csv for all files.
- Create dataframes for each wave of the study
- Drop cases (participants) from dataframes when they are no longer eligible for participation
- Add columns that identify what wave (case ID + wave)
- Extract demographics from waves and store separately from each other, with a df for demographics and a df for other survey responses for each wave.
- Export tables to sqlite format

Database Choice and Loading

We chose a SQL database due to the relational nature of the content (all being linked by the same case ID which represented an individual in the study). We settled on having separate tables for each of the following:

- Demographic data at baseline
- Demographic data at each wave
- Baseline/wave 1 data
- Wave 2 data
- Wave 3 data
- Wave 4 data
- Wave 5 data
- Wave 6 data