Capstone Project 1: Data Wrangling

Create a Google Doc (1-2 pages) describing the data wrangling steps you took to clean the dataset. Include answers to these questions in your submission:

1. What kind of cleaning steps did you perform?
    - I merged all of the datasets of interest into one large dataframe, df_full.
    - I removed values with response '99999' and '77777', which indicated responses like 'refused' and 'don't know'. I left values of NaN, which included data that was not applicable to that particular study participant.
    - Looked over data and selected columns of interest into new dataframe, df_selected
    - Viewed type of data in the dataset, number of entries, number of values in each column/series
    - Looked at the summary information for a handful of the data columns/series, to get an overview of the data and if there were any outliers. This is what led me to remove values with 99999 and 77777 responses as indicated above.

2. How did you deal with missing values, if any?

    Besides removing data as mentioned above, I left the NaN data, as this looks to be data that did not apply to those particular study participants, (and that is why those questions had the NaN response indicated).

3. Were there outliers, and how did you handle them

    There were a few outliers, but I left those in the dataset, as they overall seemed to be correct/true values and the dataset may become too small if they are removed.