Title: Motherhood in the workforce

Original Data:I have one dataset gathered from IPUMS-Highered website: https://highered.ipums.org/highered/ that was downloaded in csv format. The website also offers detailed information about what the columns represent, what the labels reflect, and a description of the number based code.
With python's pandas package, the data can be read into the system as a DataFrame. With the 'info' attribute, I found 206,703 rows and 18 columns, with a size of 28.4 MB

Spot and treat NAN data: Upon initial examination, one column contained nulls values, 'CHFAM' - Reason for changing employer or job: family related reasons. Since this question was worded specifically for family related reasons, we will assume a default answer of no. All null values will be converted to 0 and the column will be left as it is. The rest of the columns seem to have non-null values; however, after further investigation, this data has been engineered to have assigned non-null values to represent null datas. After reading the documentation, I input those values to the 'na_values' parameter of the pandas 'read_csv' function. Using .isnull() method, it shows 7 columns with null values. The weight column contained some null values, since it has no value to the study, it, along with other variables, were dropped.

The column, 'CHTOT' (number of children) contained 113521 null values. This value combined with the non-null values show that it was a result of na_values parameter. This indicates that most individuals skipped this part of the survey due to either unemployment or the question did not pertain to them. This column will be left as is until further investigation into the study.

There are six columns with the same sum of non null values (29846). They all appear to relate to the job aspects of the data. When checking the 'LFSTAT' (employment status) column, it shows the same number of entries were unemployed. This makes sense that only when a person was employed that these job related questions were recorded to be valid. These columns were left as they are.

Make entries meaningful: The following columns serves no relevance to this analysis: 'Weight', 'Sample', and 'SurID'. As a result, they were dropped..

The data set comes in an engineered format where many entries were categorized using numbers to replace real meaning. For example, the 'Gender' column has 1 and 2 to represent female and male, respectively. I used dictionaries to create a key:value pair for most of the columns. These dictionaries were reapplied to the original columns with the 'map' method.

Is the data tidy?

The definition of a tidy data from Hadley Wickham's paper states:

1. Each variable forms a column
2. Each observation forms a row
3. Each type of observation unit forms a table.

After initial observations and cleaning, the current data meets all the above requirements for tidy data. Each column and row are defined and there is only one table. This data is prepared and ready for EDA.