**Data Treatments**

* Multiple tables in a single excel sheet:  
  created Python function read\_table() to read these tables all in once. This function takes 3 arguments:
  + path: where dataset is located
  + header: list of rows where each table header starts
  + nrows: list contains number of rows for each table to be read.
* Dual header rows:

Most tables have multiple header rows. Some tables have their first header row as the correct header and some have their second!

I treated this in Python function called clean\_cols(). This function takes 3 arguments:

* + dataset: data frame name that is to be treated
  + level: row number to be considered as the correct header
  + todropna: Boolean to drop empty columns. Some of those columns result from unstacking the table header.
* Table hierarchy within table rows:

Found multiple issues here:

* + Most of the tables have category titles embedded within their rows and not in separate columns.
  + Those tables are formed as pivot tables which includes both subtotals and totals
  + Some of the titles contains spaces or tabs which couldn’t match during string comparisons
  + Some tables have the first few rows not under any category and the rest of rows categorized.

These problems were treated in a Python function called clean\_rows(). This function takes multiple inputs, below are some of them:

* + - dataset: data frame name that is to be treated
    - index\_col: Name of new column to hold the categories hierarchy names.
    - init\_index\_col\_val: to give category name to the first few rows which are not under any category.
    - cntrl\_col: Column name to investigate and decide wither current row is category or normal row.
    - valu\_col: column that contains the category name
* Column names differ from one dataset to the other:

Although these columns point to the same data, their names differ between relative datasets. Names were all united across datasets in order to give accurate analysis. Below are some examples:

'الناجحون': 'المجتازون', 'نسبة النجاح':'نسبة المجتازين

* Special characters:

Some header names or categories contain special characters like: ['\xa0', ' ــ']

which are unseen through human eyes but make a big difference when it comes to analysis as they cause same category to be analyzed as two different categories.

These were solved through Python code as well.