Data cleaning (Python) - Homework 1

Excel File - Marriage Rates

Brenden Arias

Getting a good sense of how pandas works was a good take away from this data cleaning assignment. After understanding how parts of pandas worked, it was a bit easier to finish up the code. A takeaway from this assignment is seeing just how strong this library is. In a few lines of code, you can completely reshape a set of data and write it into a manageable database.

The data structure implemented is a data frame of an excel file reshaped to a long format. Columns State, Year, and Marriage rate rearrange the data into a form more easily translatable for analysis by computers.

The approach taken to reshape the data was to stack the excel file in pandas to flatten the column headers and push the data into a long format, finally resetting the index to list rows. Then, renamed the columns to match what they are now representing and then dropping a redundant column created by stack of repeated values “Marriage Rate”. Finally, write the data frame with .to\_excel as a cleaned data file.

Furthermore, I created a database with sqlite3. Connecting to the database allow us to select data from it or create it flat out. After that we then write our data frame, marriage, using to\_sql, and create a new table of in the database of our marriage rate data. We can then create a cursor to query this table.

To use this program, enter the dirty data sheet you would like to clean into .read\_excel. Then you can set the parameters to avoid metadata. Use stack and reset\_index to reshape your data, according to its original form. You can then use rename to rename relevant columns. If there are any columns that need to be removed, you can use drop. Finally read your file into excel with to\_excel and if you would like to create a database for this cleaned data use sqlite3.

* Imported libraries
* In read to excel we started by reading in our dirty excel file. We then skipped the footer and header of metadata information with skiprows and skipfooter. We established our header, index and columns with header, usecols, and index\_col. Finally, we set null values to --- with na\_values.
* .stack().reset\_index() used to flatten column headers and reshape data.
* .rename() used to rename the new columns created by data reshaping to relevant headings.
* .drop() to get rid of redundant column with repeated value “marriage rate”
* .to\_excel() to write our cleaned and reshaped data to new clean excel file
* Sqlite3.connect() connects us or creates our database (marriage.db)
* .to\_sql() writes our data frame to a table in marriage.db
* Cursor allows us to execute queries to search our table.
* .fetchall() sets our variable results to everything we queried (fetched)