

Hamid Parsazadeh

University of toronto | stdent number:

Programming for Data Science

mid-term project

Project description:

During the past weeks you were taught how to manipulate data with Python. Different Python libraries, functions and techniques were introduced to prepare a dataset for any required analysis. “Tidy Data” was introduced as a standard approach for data cleansing. You are provided with a dataset and you are tasked to restructure this dataset based on the Tidy Data by Python. The final dataset should be fully aligned with the Tidy Data principles. You should submit three files at the end of the project: Cleaned Dataset, Python Codes used to clean this dataset and a short description of the steps were taken in this project.

1. **Dataset Analysis:**

The provided dataset is a summary report provided by Department of Economic and Social Affairs of the United Nations (UN). This dataset reports on Trends in International Migrant Stock as the 2015 Revision. It includes six data tables and two descriptive tables. The general structure of this dataset is as follow:

**Table 1**

**Table 2**

**Table 4**

**Table 5**

**Table 6**

**Table 3**

Tables one to six connect with the Annex by “Country or area” or “Country Code” and connect with the Note based on an attribute which we can name it as “Note”. There are a few levels of aggregation in the tables which are managed by fields: “Region” and “Major Area” and attributes: “Developed region”, “Least developed country” and “Sub-Saharan Africa”. The fields “Region” and “Major Area” have exclusive “Code” and “Sort Order”.

* 1. **Key observations:**

The Annex table is slightly defective as it does not represent the Codes are assigned to “World” as 900, “Developed regions” as 901, “Developing regions” as 902, “Least developed countries” as 941 and “Less developed regions excluding least developed countries” as 934. The Sort order for these categories is also missed in Annex

Tables: one and two summaries data for both sexes: “Male and “Female” and also the “Total”. It seems that there is no “unisex” or “unknown” category in this dataset. As a result, the section “Total” is fully calculated.

Table three is fully calculated as “percentage of the total population by sex and by major area, region, country or area, 1990-2015” from Table 1 and Table 2 [1]

Table four is are also fully calculated as “percentage of the international migrant stock by major area, region, country or area, 1990-2015” from Table 1 [1]

Table 5 is a calculated table as “Annual rate of change of the migrant stock by sex and by major area, region, country or area, 1990-2015 (percentage)” but it is not fully calculated from the given data in this dataset.

Table six includes three sections:

* Section 1: “Estimated refugee stock at mid-year by major area, region, country or area, 1990-2015”
  + It seems to be the estimation of total refugees without gender breakdown
* Section 2: Refugee as a percentage of international migrant stock
  + It is a calculated section based on section 1 and Table 1
* Section 3: Annual rate of change of the refugee stock
  + It is a calculated section but not fully supported by the available data in this dataset

Finally, numbers in Table 2 are rounded to thousand. As a result, while we can reverse this rounding by multiplying them with 1000, it won’t be accurate as we have no access the lost information.

1. **Dataset import to Python**

Using Python functions and Pandas library the dataset is broken down to seven csv files: Table 1 to six, Annex, Note and Contents to ensure, without missing any information, the sections and tables of this dataset are ready to be modified based on the Tidy Data principles.

Column names are adjusted to be unique and relevant. For instance, the original dataset has three variables named as “Sort order”, which are sorting values based on different objectives. At this stage, these variables are assigned modified names so, they will be unique and related to their objective.

1. **Dataset Cleansing based on Tidy Data**

* **Adding the missed codes to Annex**
* **Dropping the aggregated / calculated rows**
* **Dropping the aggregated / calculated columns**
* **Melting tables by adding year and Gender**
* **Fixing the rounding issue in Table 2 (multiplying its value to 1000)**
* **Merging the new Table 1 and 2 and maybe 3 by adding a new attribute as their identifier**

References:

1. *Trends in international migrant stock: The 2015 revision - un*. (n.d.). Retrieved November 2, 2021, from <https://www.un.org/en/development/desa/population/migration/data/estimates2/docs/MigrationStockDocumentation_2015.pdf>