ENSF 592 Spring 2021 – Final Project Report

Authors: Bhavyai Gupta, Brandon Attai
Date: 2021-06-15
Course: ENSF 592

Summary

This program allows a user to request, select or visualize various statistics surrounding United Nations Data to give insight into population and wealth trends for UN Regions, Sub-Regions and Countries. To achieve this, a total of four Datasets were used as follows:

1. UN Region, Sub-Region and Country
2. Population Growth, Fertility, Life Expectancy and Mortality
3. Population in the Capital City, Urban and Rural Areas
4. GDP and GDP per Capita

The program makes use of four Python files to provide an interactive command line user interface. The interface makes use of color-coded messages to allow for ease of readability. Once the user runs the launch.python file that contains main, the user is given a real time update on the steps being completed and their status via the terminal. After each step is completed, the user is updated with the status. The user is informed of information such as, the data being merged into one DataFrame and when the check for null values is completed.

The user is then prompted to enter the program menu. Once entered, the user has 7 options to select from, which allows for exporting, requesting, selecting, or visualizing the statistics. For example, if the user wishes print the aggregate statistics for GDP per capita with respect to the USA or Ratio of Urban Population to GDP per Capita amongst others, the user can select option [4] and print the aggregate statistics for a Region or Sub-Region based on the user's text entry. Table 1 below summarizes the other functionality within the program.

Item	Options
1	Print the imported datasets
2	Re-export the entire merged hierarchical dataset into Excel
3	Print aggregate stats for the entire dataset
4	Print aggregation stats grouped by UN Region/UN Sub-Region and available years
5	Print the list of countries that have higher GDP per capita than USA, and the year
6	Show plot of Population Increase, Total Fertility Rate and Life Expectancy for a country
0	Exit

The program makes use of object-oriented programming by using classes and methods to handle data analysis and calculations. Finally, exception handling is performed throughout the program to ensure the program does not terminate if the user enters an invalid input.

A matrix summarizing how the requirements are met are shown in Appendix 1 on page 3.

References

1. UN Region, Sub-Region and Country, Development Data Section of the Development Data and Outreach Branch within the Statistics Division of the Department of Economic and Social Affairs (UN DESA) of the UN Secretariat, June 2019. [Online]. Available: https://data.un.org/_Docs/SYB/CSV/SYB63_1_202105_Population,%20Surface%20Area%20and%20Density.csv

2. Population Growth, Fertility, Life Expectancy and Mortality, Development Data Section of the Development Data and Outreach Branch within the Statistics Division of the Department of Economic and Social Affairs (UN DESA) of the UN Secretariat, Aug. 2019. [Online]. Available: https://data.un.org/_Docs/SYB/CSV/SYB62_246_201907_Population%20Growth,%20Fertility%20and%20Mortality%20Indicators.csv

3. Population in the Capital City, Urban and Rural Areas, Development Data Section of the Development Data and Outreach Branch within the Statistics Division of the Department of Economic and Social Affairs (UN DESA) of the UN Secretariat, May 2018. [Online]. Available: https://data.un.org/_Docs/SYB/CSV/SYB61_253_Population%20Growth%20Rates%20in%20Urban%20areas%20and%20Capital%20cities.csv

4. GDP and GDP per Capita, Development Data Section of the Development Data and Outreach Branch within the Statistics Division of the Department of Economic and Social Affairs (UN DESA) of the UN Secretariat, Nov. 2020. [Online]. Available: https://data.un.org/_Docs/SYB/CSV/SYB63_230_202009_GDP%20and%20GDP%20Per%20Capita.csv

Appendix 1

Item	Stages	Objective	Notes
1	Stage 1: Dataset Selection	Several suggested datasets are included in the project repository. You may use the provide data or select datasets of your own choosing.	See Github Repository folders - UN Custom Data & UN Population Datasets.
2		You must use at least three separate Excel sheets or files that can be related in some way.	Four separate Excel sheets were used.
3		Your final combined dataset (see next stage) must have at least ten columns and 200 rows.	12 Columns inclusive of indices.
4		You may edit the given datasets before you begin coding, but your program should not modify the Excel files directly.	Program does not modify the excel files directly.
5		You may not hard-code/copy-paste any information into your program except for the Excel column names.	No information hard-code or copy-paste within the program except column names.
6	Stage 2: DataFrame Creation	Import your chosen data into a Pandas DataFrames.	Import is done using "import_data" method in class DataAnalysis.
7		You must use at least two merge/join operations and you must delete any duplicated columns/rows that result from the merge.	Merge is done using "merge_data" method in class Data Analysis.
8		You must create a hierarchical index of at least two levels (row or column).	Two level column row index is created in "merge_data" method (line 189) in class DataAnalysis.
9		All data should be presented in the correctly sorted order, depending on the index.	Data is sorted in "merge_data" method (line 192) in class DataAnalysis.
11		You may not use global variables. You must import the data within your main function.	No global variables used.
12		Remember to check for null values or data mismatches.	Created "check_null" method (line 258)for this. This is just a formality as dropna() is used.
13	Stage 3: User Entry	Your application must return useful information. Design an interface that allows users to search based on some sort of criteria or keywords.	User can search for information by selection and keywords/criteria.
14		The user must provide at least two pieces of information/selection (e.g. "school name" and "grade")	Option [4] in the user interface allows for two entries to customize aggregate data.
15		Give the user clear input instructions. If an invalid entry is given, use try/except statements to handle the error. Your program should not terminate.	Try/Except statements used to handle invalid user entries throughout.
16		You must not hard-code any data values (the data within your spreadsheets could be changed!).	Reference item 5.
17		Any output information must be clearly defined using printed headers.	Printed headers and color coding utilized.
18	Stage 4: Analysis and Calculations	You may choose what data trends to presents from your data. However, you must meet the following specifications.	
19		Use the describe method to print aggregate stats for the entire dataset.	Describe method used in "print_aggregate_stats" method (line 314) in DataAnalysis class.
20		Add at least two columns to the combined dataset.	Additional columns added in "additional_statistics" method (line 197) in DataAnalysis class.
21		Use an aggregation computation for a subset of the data.	Aggregation operation used in "group_by_stats" method (line 328) in DataAnalysis class.
22		Use a masking operation.	Masking operation used in "higher_gdp_than_usa" method (line 418) in DataAnalysis class.
23		Use the groupby operation at least once.	Groupby operation used in "group_by_stats" method (line 328) in DataAnalysis class.
24		Create and print a pivot table.	Pivot Table created (line 508) and printed (line 513) in "pivot_plot" method in DataAnalysis class.
25		Include at least two user-defined functions or a class that contains two methods.	More than two functions/methods used.
27	Stage 5: Export and Matplotlib	Export your entire merged, hierarchical dataset to an Excel file in the working directory. Be sure to include the index and header values. The TAs will use this to verify the structure of your dataset.	Data is exported using "export_dataset" method (line270) in class DataAnalysis.
28		Use your data to create at least one plot using Matplotlib. Save the plot as a .png file and upload to the repository.	Plots are created using Matplotlib in "pivot_plot" method (line 517 onwards) in class DataAnalysis.