**ENSF 592 Programming Fundamentals for Data Engineers**

**Final Project**

**The Batt Boys – Group 21**

**Nathan Tham 30046119**

**Justin Nguyen 30042258**

[**https://github.com/MaanKhedr-teaching/final-project-the-batt-boys.git**](https://github.com/MaanKhedr-teaching/final-project-the-batt-boys.git)

**Introduction**

For the final project of ENSF 592 a terminal-based data analysis program was created in Python. The program followed the following design steps: Dataset selection, Data Frame Creation, User Entry, Analysis and Calculations, and Export and Matplotlib. The dataset that we used was the UN Population Datasets.

**Task Distribution**

Justin: Data frame creation and user prompt logic.

Nathan: Analysis, calculations, and matplotlib plot.

Live Share extension was used to collaborate throughout the project.

**Program Breakdown**

Stage 1: Dataset Selection

Datasets used were UN Codes, UN Population Dataset 1, and UN Population Dataset 2 which were all in .xlxs format. UN Population Dataset 1 and 2 were concatenated together to form one large UN Population Dataset, and then was sorted by Code. The concatenated UN Population Dataset was separated between an index and a value data frame for ease in creating the combined data frame in the next stage. The final combined data frame has 6000+ rows and 11 columns.

Stage 2: Data Frame Creation

A combined multi-index data frame was created which included all the data from all the files. The rows were 3 level indexed as Type -> Region/Country/Area -> Code. The columns were 1 level indexed with headers Index, Year, Series, Capital City, Value, and Dataset (*created an additional 5 columns later*). The combined dataset was sorted by Code and exported as combined\_df.xlxs in the project directory.

Stage 3: User Entry

The Region/Country/Area and their associated codes were placed into two lists. The program then enters a loop, prompting user input for Region/Country/Area or Code within try and except statements. The user input is checked against the Region/Country/Area and codes list for validity. If the input is invalid, there is exception handling which allows the loop to prompt the user again without exiting the program.

A new filtered data frame is created based on the user input where the data is sorted and sliced. A pivot table and chart are plotted for this data and saved as figure.png.

The user is once again prompted to choose between data from the first or second dataset. Similar to before, this statement is nested within a loop, try, and except statements which allows for proper error handling. Once the user input has been confirmed as valid, the filtered data frame is updated with more precise information. The requested dataset is isolated using masking.

Stage 4: Analysis and Calculations

At the beginning of the program, the aggregate statistics are calculated for the entire dataset using the describe() method. Statistics are then calculated for the requested dataset. Mean, standard deviation, minimum value, maximum value, total sum is all calculated for the requested data and added to the final excel dataset. The statistics are calculated using the groupby() method being passed the ‘Series’ column. A pivot table is then created which calculates and displays the same values that will be placed into the excel file. Finally, the values are returned and written into the excel file.

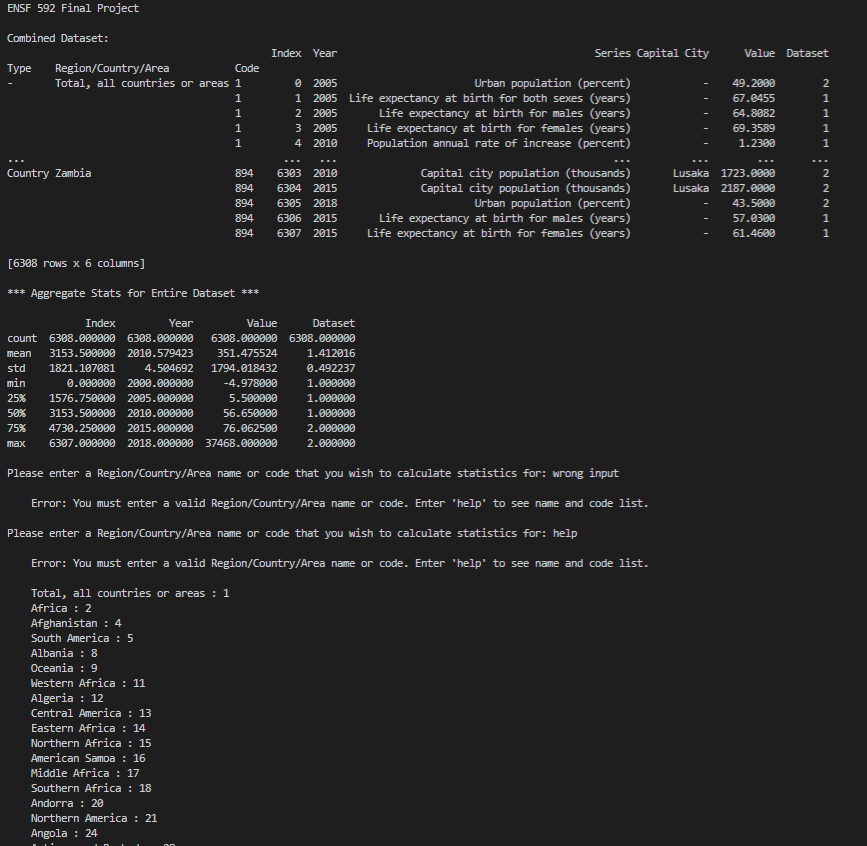
Stage 5: Export and Matplotlib

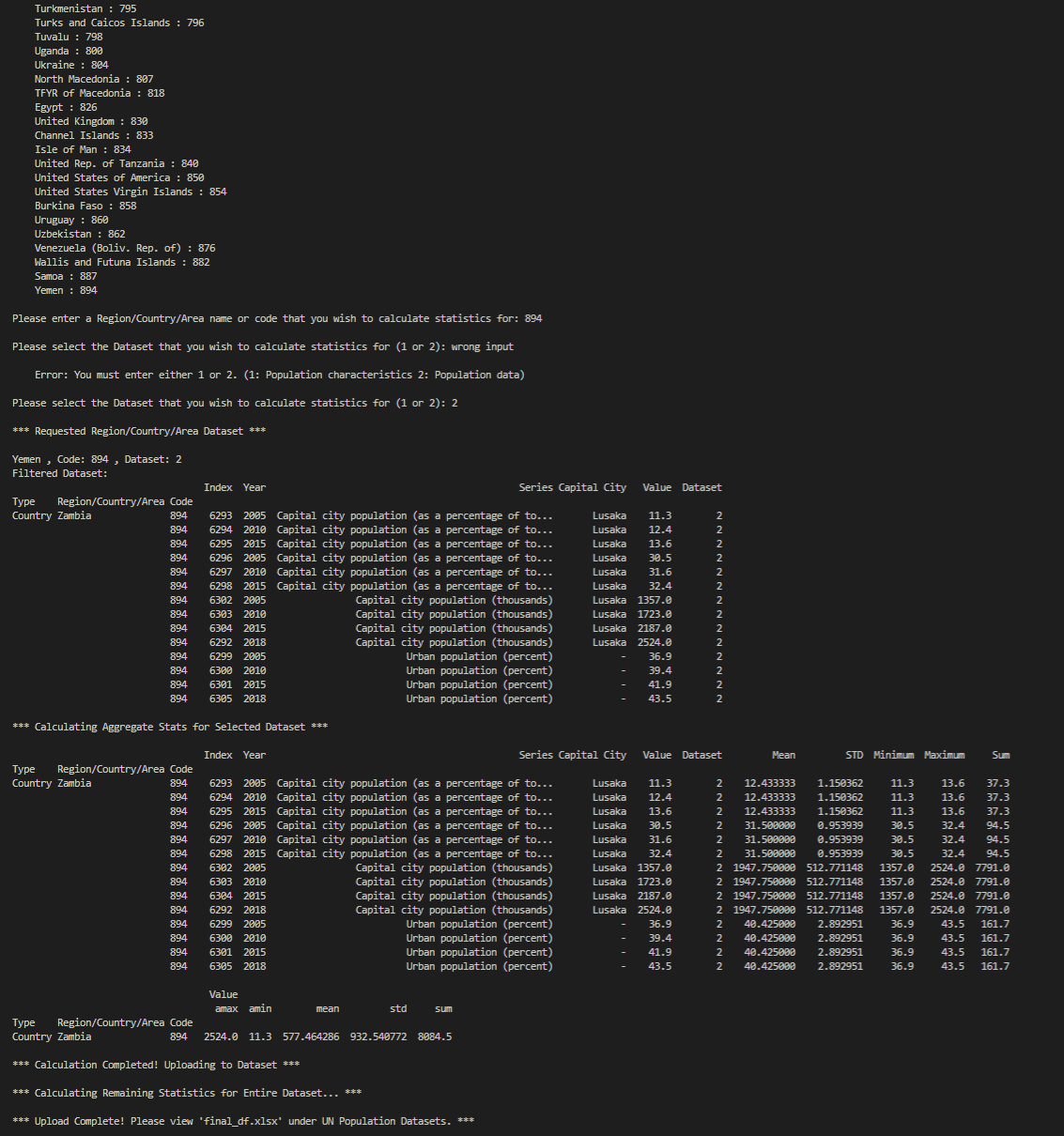
Since the calculations and data frame creation was completed in the previous section, this section solely consists of exporting the final data frame as a .xlxs file.

**Screenshots**

Text

Description automatically generated





References:

[1] “SYB62\_246\_201907\_Population Growth, Fertility and Mortality Indicators”, United Nations. [Online]. Available: http://data.un.org/. [Accessed: 13-Jun-2022].

[2] “SYB61\_253\_Population Growth Rates in Urban areas and Capital cities”, United Nations. [Online]. Available: http://data.un.org/. [Accessed: 13-Jun-2022].

[3] “UN Population Datasets”, Gapminder. [Online]. Available: https://www.gapminder.org/data/. [Accessed: 14-Jun-2022].

[4] “UN Population Datasets”, United Nations. [Online]. Available: https://unstats.un.org/unsd/methodology/m49/ [Accessed: 14-Jun-2022].