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

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.

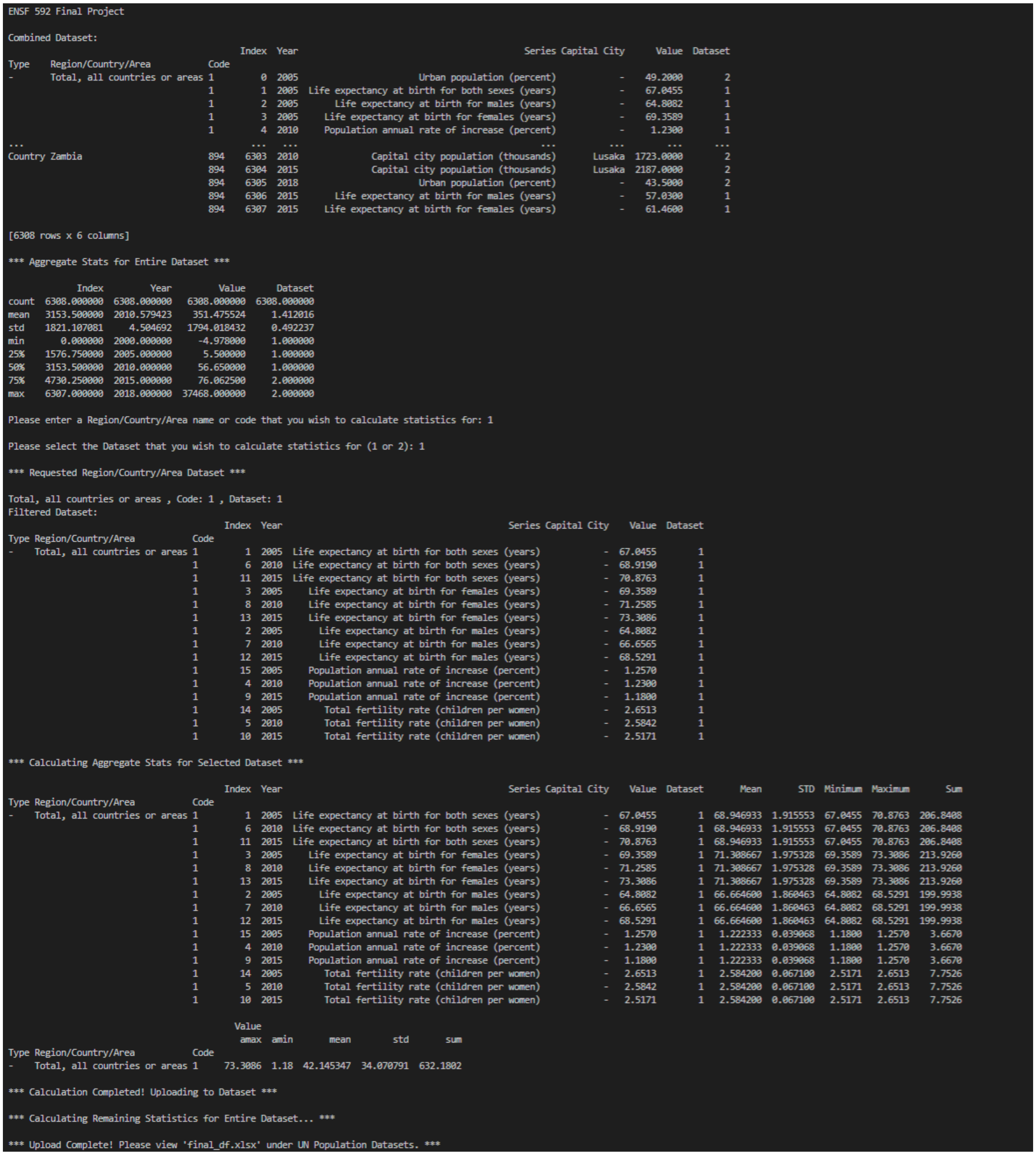
Stage 4: Analysis and Calculations

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.

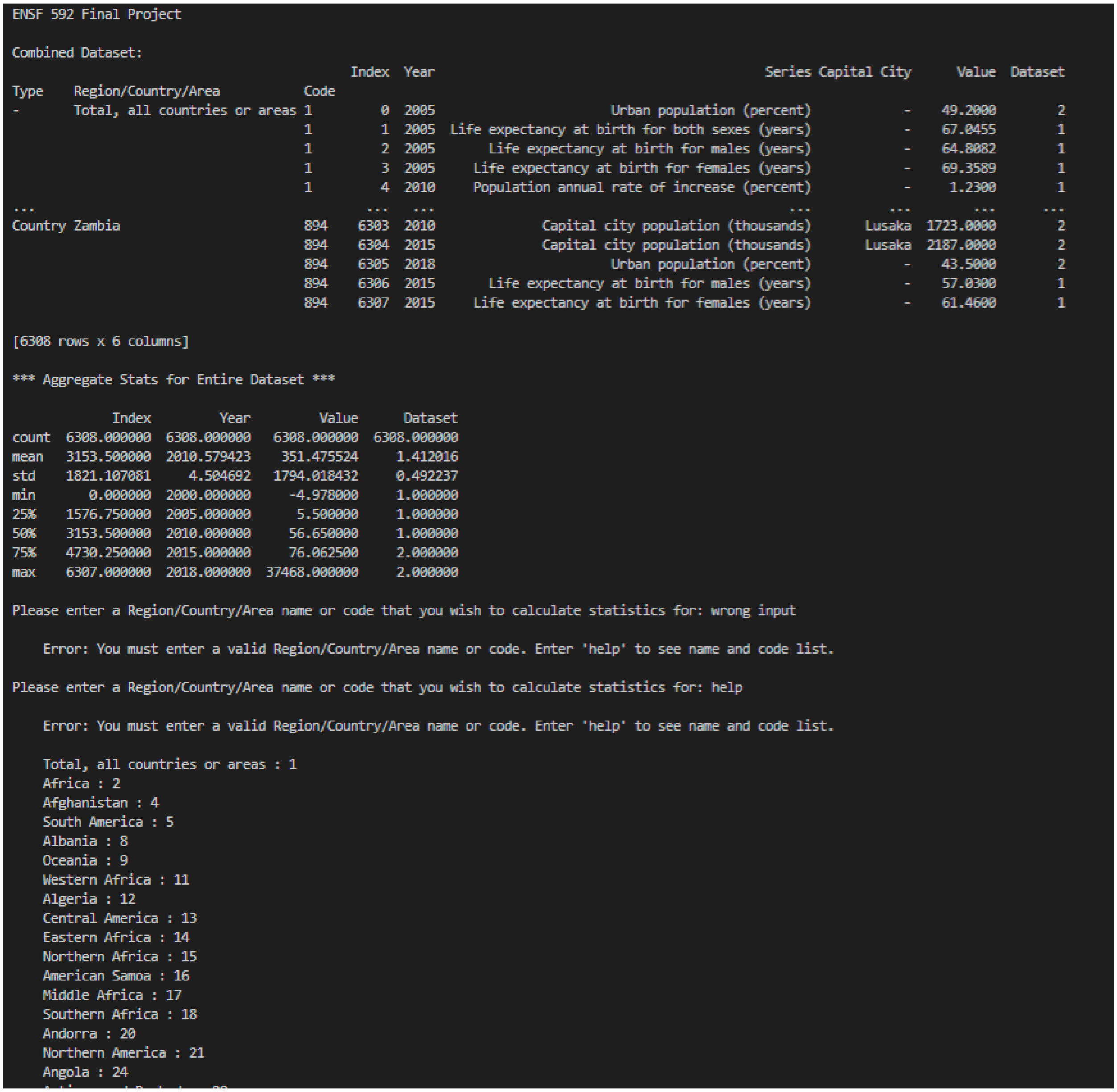
Stage 5: Export and Matplotlib

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.

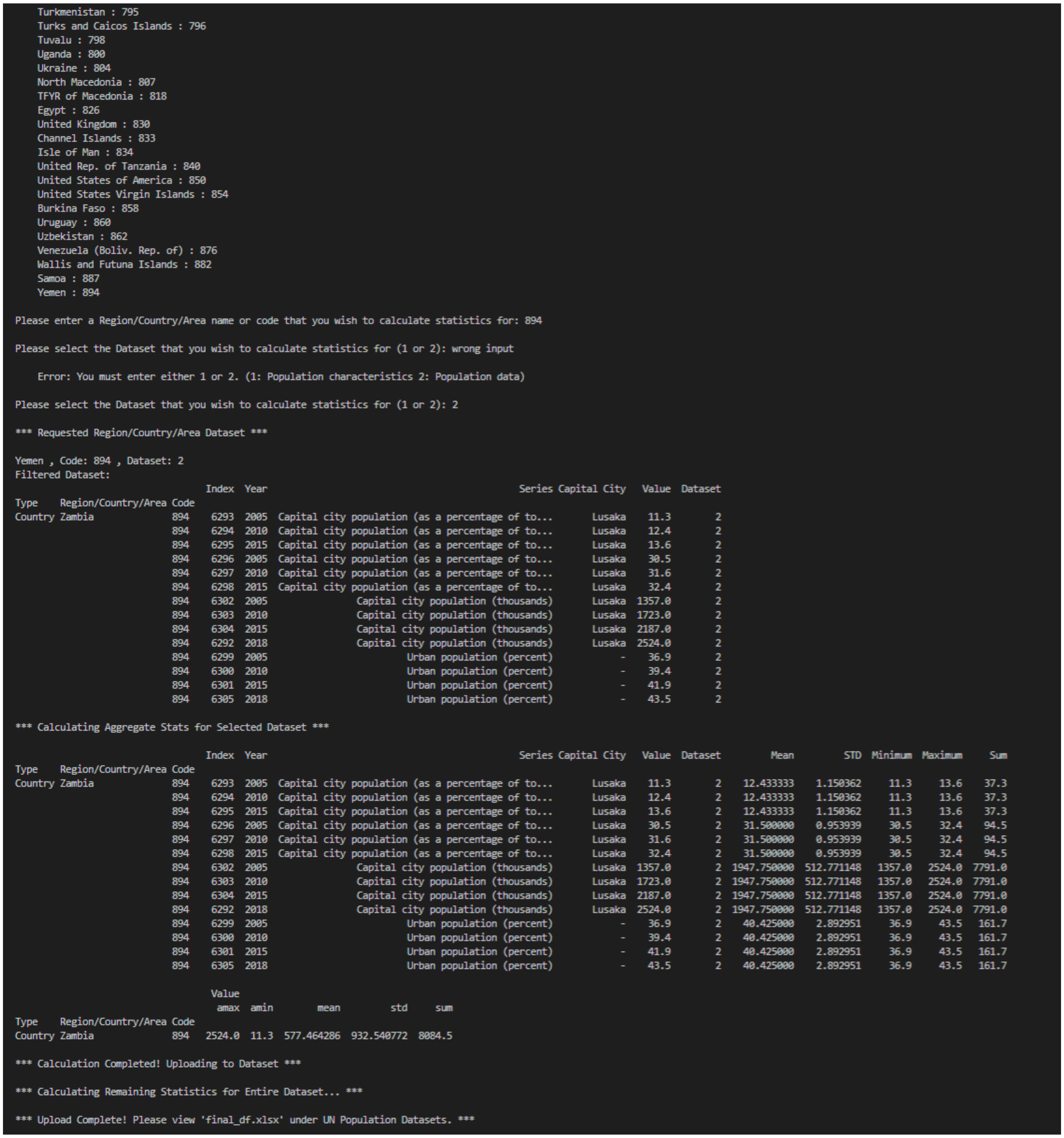
Screenshots



```
ENSF 592 Final Project

Combined Dataset:
                                            Index  Year                                           Series  Capital City      Value  Dataset
Type    Region/Country/Area           Code
-       Total, all countries or areas 1         0  2005                       Urban population (percent)             -    49.2000        2
                                      1         1  2005  Life expectancy at birth for both sexes (years)             -    67.0455        1
                                      1         2  2005       Life expectancy at birth for males (years)             -    64.8082        1
                                      1         3  2005     Life expectancy at birth for females (years)             -    69.3589        1
                                      1         4  2010     Population annual rate of increase (percent)             -     1.2300        1
...                                           ...   ...                                              ...           ...        ...      ...
Country Zambia                        894    6303  2010              Capital city population (thousands)        Lusaka  1723.0000        2
                                      894    6304  2015              Capital city population (thousands)        Lusaka  2187.0000        2
                                      894    6305  2018                       Urban population (percent)             -    43.5000        2
                                      894    6306  2015       Life expectancy at birth for males (years)             -    57.0300        1
                                      894    6307  2015     Life expectancy at birth for females (years)             -    61.4600        1

[6308 rows x 6 columns]

*** Aggregate Stats for Entire Dataset ***

             Index         Year         Value      Dataset
count  6308.000000  6308.000000   6308.000000  6308.000000
mean   3153.500000  2010.579423    351.475524     1.412016
std    1821.107081     4.504692   1794.018432     0.492237
min       0.000000  2000.000000     -4.978000     1.000000
25%    1576.750000  2005.000000      5.500000     1.000000
50%    3153.500000  2010.000000     56.650000     1.000000
75%    4730.250000  2015.000000     76.062500     2.000000
max    6307.000000  2018.000000  37468.000000     2.000000

Please enter a Region/Country/Area name or code that you wish to calculate statistics for: 1

Please select the Dataset that you wish to calculate statistics for (1 or 2): 1

*** Requested Region/Country/Area Dataset ***

Total, all countries or areas , Code: 1 , Dataset: 1
Filtered Dataset:
                                         Index  Year                                           Series  Capital City    Value  Dataset
Type Region/Country/Area           Code
-    Total, all countries or areas 1         1  2005  Life expectancy at birth for both sexes (years)             -  67.0455        1
                                   1         6  2010  Life expectancy at birth for both sexes (years)             -  68.9190        1
                                   1        11  2015  Life expectancy at birth for both sexes (years)             -  70.8763        1
                                   1         3  2005     Life expectancy at birth for females (years)             -  69.3589        1
                                   1         8  2010     Life expectancy at birth for females (years)             -  71.2585        1
                                   1        13  2015     Life expectancy at birth for females (years)             -  73.3086        1
                                   1         2  2005       Life expectancy at birth for males (years)             -  64.8082        1
                                   1         7  2010       Life expectancy at birth for males (years)             -  66.6565        1
                                   1        12  2015       Life expectancy at birth for males (years)             -  68.5291        1
                                   1        15  2005     Population annual rate of increase (percent)             -   1.2570        1
                                   1         4  2010     Population annual rate of increase (percent)             -   1.2300        1
                                   1         9  2015     Population annual rate of increase (percent)             -   1.1800        1
                                   1        14  2005        Total fertility rate (children per women)             -   2.6513        1
                                   1         5  2010        Total fertility rate (children per women)             -   2.5842        1
                                   1        10  2015        Total fertility rate (children per women)             -   2.5171        1

*** Calculating Aggregate Stats for Selected Dataset ***

                                         Index  Year                                           Series  Capital City    Value  Dataset       Mean       STD  Minimum  Maximum       Sum
Type Region/Country/Area           Code
-    Total, all countries or areas 1         1  2005  Life expectancy at birth for both sexes (years)             -  67.0455        1  68.946933  1.915553  67.0455  70.8763  206.8408
                                   1         6  2010  Life expectancy at birth for both sexes (years)             -  68.9190        1  68.946933  1.915553  67.0455  70.8763  206.8408
                                   1        11  2015  Life expectancy at birth for both sexes (years)             -  70.8763        1  68.946933  1.915553  67.0455  70.8763  206.8408
                                   1         3  2005     Life expectancy at birth for females (years)             -  69.3589        1  71.308667  1.975328  69.3589  73.3086  213.9260
                                   1         8  2010     Life expectancy at birth for females (years)             -  71.2585        1  71.308667  1.975328  69.3589  73.3086  213.9260
                                   1        13  2015     Life expectancy at birth for females (years)             -  73.3086        1  71.308667  1.975328  69.3589  73.3086  213.9260
                                   1         2  2005       Life expectancy at birth for males (years)             -  64.8082        1  66.664600  1.860463  64.8082  68.5291  199.9938
                                   1         7  2010       Life expectancy at birth for males (years)             -  66.6565        1  66.664600  1.860463  64.8082  68.5291  199.9938
                                   1        12  2015       Life expectancy at birth for males (years)             -  68.5291        1  66.664600  1.860463  64.8082  68.5291  199.9938
                                   1        15  2005     Population annual rate of increase (percent)             -   1.2570        1   1.222333  0.039068   1.1800   1.2570    3.6670
                                   1         4  2010     Population annual rate of increase (percent)             -   1.2300        1   1.222333  0.039068   1.1800   1.2570    3.6670
                                   1         9  2015     Population annual rate of increase (percent)             -   1.1800        1   1.222333  0.039068   1.1800   1.2570    3.6670
                                   1        14  2005        Total fertility rate (children per women)             -   2.6513        1   2.584200  0.067100   2.5171   2.6513    7.7526
                                   1         5  2010        Total fertility rate (children per women)             -   2.5842        1   2.584200  0.067100   2.5171   2.6513    7.7526
                                   1        10  2015        Total fertility rate (children per women)             -   2.5171        1   2.584200  0.067100   2.5171   2.6513    7.7526

                                           Value
                                            amax  amin       mean        std       sum
Type Region/Country/Area           Code
-    Total, all countries or areas 1     73.3086  1.18  42.145347  34.070791  632.1802

*** Calculation Completed! Uploading to Dataset ***

*** Calculating Remaining Statistics for Entire Dataset... ***

*** Upload Complete! Please view 'final_df.xlsx' under UN Population Datasets. ***
```



```
ENSF 592 Final Project

Combined Dataset:
                                            Index  Year                                             Series Capital City      Value  Dataset
Type    Region/Country/Area           Code
-       Total, all countries or areas 1         0  2005                       Urban population (percent)            -    49.2000        2
                                      1         1  2005  Life expectancy at birth for both sexes (years)            -    67.0455        1
                                      1         2  2005       Life expectancy at birth for males (years)            -    64.8082        1
                                      1         3  2005     Life expectancy at birth for females (years)            -    69.3589        1
                                      1         4  2010     Population annual rate of increase (percent)            -     1.2300        1
...                                           ...   ...                                              ...          ...        ...      ...
Country Zambia                        894    6303  2010              Capital city population (thousands)       Lusaka  1723.0000        2
                                      894    6304  2015              Capital city population (thousands)       Lusaka  2187.0000        2
                                      894    6305  2018                       Urban population (percent)            -    43.5000        2
                                      894    6306  2015       Life expectancy at birth for males (years)            -    57.0300        1
                                      894    6307  2015     Life expectancy at birth for females (years)            -    61.4600        1

[6308 rows x 6 columns]

*** Aggregate Stats for Entire Dataset ***

             Index         Year         Value      Dataset
count  6308.000000  6308.000000   6308.000000  6308.000000
mean   3153.500000  2010.579423    351.475524     1.412016
std    1821.107081     4.504692   1794.018432     0.492237
min       0.000000  2000.000000     -4.978000     1.000000
25%    1576.750000  2005.000000      5.500000     1.000000
50%    3153.500000  2010.000000     56.650000     1.000000
75%    4730.250000  2015.000000     76.062500     2.000000
max    6307.000000  2018.000000  37468.000000     2.000000

Please enter a Region/Country/Area name or code that you wish to calculate statistics for: wrong input

    Error: You must enter a valid Region/Country/Area name or code. Enter 'help' to see name and code list.

Please enter a Region/Country/Area name or code that you wish to calculate statistics for: help

    Error: You must enter a valid Region/Country/Area name or code. Enter 'help' to see name and code list.

    Total, all countries or areas : 1
    Africa : 2
    Afghanistan : 4
    South America : 5
    Albania : 8
    Oceania : 9
    Western Africa : 11
    Algeria : 12
    Central America : 13
    Eastern Africa : 14
    Northern Africa : 15
    American Samoa : 16
    Middle Africa : 17
    Southern Africa : 18
    Andorra : 20
    Northern America : 21
    Angola : 24
```



```
    Turkmenistan : 795
    Turks and Caicos Islands : 796
    Tuvalu : 798
    Uganda : 800
    Ukraine : 804
    North Macedonia : 807
    TFYR of Macedonia : 818
    Egypt : 826
    United Kingdom : 830
    Channel Islands : 833
    Isle of Man : 834
    United Rep. of Tanzania : 840
    United States of America : 850
    United States Virgin Islands : 854
    Burkina Faso : 858
    Uruguay : 860
    Uzbekistan : 862
    Venezuela (Boliv. Rep. of) : 876
    Wallis and Futuna Islands : 882
    Samoa : 887
    Yemen : 894

Please enter a Region/Country/Area name or code that you wish to calculate statistics for: 894

Please select the Dataset that you wish to calculate statistics for (1 or 2): wrong input

    Error: You must enter either 1 or 2. (1: Population characteristics 2: Population data)

Please select the Dataset that you wish to calculate statistics for (1 or 2): 2

*** Requested Region/Country/Area Dataset ***

Yemen , Code: 894 , Dataset: 2
Filtered Dataset:
                                  Index  Year                                             Series Capital City   Value  Dataset
Type    Region/Country/Area Code
Country Zambia              894    6293  2005  Capital city population (as a percentage of to...       Lusaka    11.3        2
                            894    6294  2010  Capital city population (as a percentage of to...       Lusaka    12.4        2
                            894    6295  2015  Capital city population (as a percentage of to...       Lusaka    13.6        2
                            894    6296  2005  Capital city population (as a percentage of to...       Lusaka    30.5        2
                            894    6297  2010  Capital city population (as a percentage of to...       Lusaka    31.6        2
                            894    6298  2015  Capital city population (as a percentage of to...       Lusaka    32.4        2
                            894    6302  2005                Capital city population (thousands)       Lusaka  1357.0        2
                            894    6303  2010                Capital city population (thousands)       Lusaka  1723.0        2
                            894    6304  2015                Capital city population (thousands)       Lusaka  2187.0        2
                            894    6292  2018                Capital city population (thousands)       Lusaka  2524.0        2
                            894    6299  2005                         Urban population (percent)            -    36.9        2
                            894    6300  2010                         Urban population (percent)            -    39.4        2
                            894    6301  2015                         Urban population (percent)            -    41.9        2
                            894    6305  2018                         Urban population (percent)            -    43.5        2

*** Calculating Aggregate Stats for Selected Dataset ***

                                  Index  Year                                             Series Capital City   Value  Dataset         Mean         STD  Minimum  Maximum     Sum
Type    Region/Country/Area Code
Country Zambia              894    6293  2005  Capital city population (as a percentage of to...       Lusaka    11.3        2    12.433333    1.150362     11.3     13.6    37.3
                            894    6294  2010  Capital city population (as a percentage of to...       Lusaka    12.4        2    12.433333    1.150362     11.3     13.6    37.3
                            894    6295  2015  Capital city population (as a percentage of to...       Lusaka    13.6        2    12.433333    1.150362     11.3     13.6    37.3
                            894    6296  2005  Capital city population (as a percentage of to...       Lusaka    30.5        2    31.500000    0.953939     30.5     32.4    94.5
                            894    6297  2010  Capital city population (as a percentage of to...       Lusaka    31.6        2    31.500000    0.953939     30.5     32.4    94.5
                            894    6298  2015  Capital city population (as a percentage of to...       Lusaka    32.4        2    31.500000    0.953939     30.5     32.4    94.5
                            894    6302  2005                Capital city population (thousands)       Lusaka  1357.0        2  1947.750000  512.771148   1357.0   2524.0  7791.0
                            894    6303  2010                Capital city population (thousands)       Lusaka  1723.0        2  1947.750000  512.771148   1357.0   2524.0  7791.0
                            894    6304  2015                Capital city population (thousands)       Lusaka  2187.0        2  1947.750000  512.771148   1357.0   2524.0  7791.0
                            894    6292  2018                Capital city population (thousands)       Lusaka  2524.0        2  1947.750000  512.771148   1357.0   2524.0  7791.0
                            894    6299  2005                         Urban population (percent)            -    36.9        2    40.425000    2.892951     36.9     43.5   161.7
                            894    6300  2010                         Urban population (percent)            -    39.4        2    40.425000    2.892951     36.9     43.5   161.7
                            894    6301  2015                         Urban population (percent)            -    41.9        2    40.425000    2.892951     36.9     43.5   161.7
                            894    6305  2018                         Urban population (percent)            -    43.5        2    40.425000    2.892951     36.9     43.5   161.7

                                   Value
                                    amax  amin        mean         std     sum
Type    Region/Country/Area Code
Country Zambia              894   2524.0  11.3  577.464286  932.540772  8084.5

*** Calculation Completed! Uploading to Dataset ***

*** Calculating Remaining Statistics for Entire Dataset... ***

*** Upload Complete! Please view 'final_df.xlsx' under UN Population Datasets. ***
```

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].