

# ENSF 692: Final Project (Spring 2025)

Authors: Conner Castle & William Watson

## 1 Introduction

This report describes the functionality of the accompanying python file 'data\_analysis.py'. The purpose is to:

1. Document the data processing workflow used to transform three separate sets of macroeconomic data in csv format into a unified data frame that has undergone several processing and analysis steps and eventually been converted back to a new csv file.
2. Describe the command line interface functionality that allows for user specified analysis of subsets of the data.
3. Explain how and where the project requirements have been met. These will be referenced to either function names or processing steps in main().

Ancillary files such as Jupyter notebooks are included in the GitHub repository for reference, as these were the primary method for exploratory analysis and processing.

## 2 Datasets

This investigation focuses on the relationship between the GDP of nations and their levels of domestic income inequality. The observation that motivated this investigation is the following: Ireland currently has the highest GDP Per Capita in the world. However, it is widely known that this is not the result of an increasingly powerful economy. Rather, it is due to the massive influx of direct foreign investment (DFI) because of a sharp reduction in national corporate tax rate. Ireland has become a tax haven for multinationals operating in Europe. The spike in GDP (and by extension, GDP Per Capita) has been labeled an accounting trick, and satirically dubbed 'Leprechaun Economics'. So, how does this spike in GDP affect the quality of life for the inhabitants of Ireland? Is it a merely a vapid statistic, or is there a tangible benefit to the individual people?

Data Sources:

Total GDP (US\$, inflation-adjusted), GapMinder, June 2025 [Online]

<https://www.gapminder.org/data/>

GDP Per Capita (US\$, inflation-adjusted), GapMinder, June 2025 [Online]

<https://www.gapminder.org/data/>

Gini coefficient, GapMinder, June 2025 [Online] <https://www.gapminder.org/data/>

Note: The Gini Coefficient is a measure of national income inequality, represented as a value between zero and 100 where zero is perfect equality, and 100 in maximum inequality.

### 3 Data Processing Workflow

The data was processed in the following steps:

#### 3.1 Data Cleaning and Merging

1. Read in raw data from CSV files to Pandas data frames ('Total GDP', 'GDP Per Capita', and 'Gini Coeff' data.)
2. Sort and copy the raw data to perform manipulation without modifying the original data.
3. Parse the Total GDP and GDP Per Capita data frame values to convert from strings to floats.
  - E.g., '1.78TN' → 1,780,000,000,000.0
4. Merge the datasets using an inner join on the rows and columns. This ensures only common countries and years are preserved. The combined dataset is MultiIndexed on the rows, where the outer index is the 'Metric': ['Total GDP', 'GDP Per Capita', 'Gini Coeff'], and the inner index is the 'Country'.
5. Filter rows (countries) based on the number of nan values. In this case, a country is dropped if less than 50% of years contained real values in either 'Total GDP' or 'GDP Per Capita'.

#### 3.2 General Aggregate Statistics

6. Show aggregate statistics for the cleaned datasets using the *describe()* method applied to each metric.

#### 3.3 Derive New Data

7. Create new dataset from existing data to reflect the GDP Per Capita adjusted for inequality. To do this, a new metric called "Gini Dollars", which is defined and calculated as follows:

$$\text{Gini Dollars} = \text{GDP Per Capita} * (1 - \text{Gini Coeff})$$

Where *Gini Coeff* is a value between (0 – 1).

This metric for each year and country is then added as rows to the data frame under the new metric index '*Gini Dollars*'.

### 3.4 Add Level to MultiIndex Rows

8. Create a new classification label for each country based on the 'Total GDP' metric, the classes are defined as follows:
  - 'Superpower' : > 10 trillion USD
  - 'Heavyweight' : 1 trillion – 10 trillion USD
  - 'Middleweight' : 100 billion – 1 trillion USD
  - 'Lightweight' : < 100 billion USD
9. Use the *describe()* method again to show aggregate stats for all data, including the newly derived additional data.
10. Show a pivot table of averaged values for a specific year (default is end year of user input range), pivoting on 'Metric' and 'Classification'.

### 3.5 User Input

11. Prompt the user to input a range for the years to include in the analysis. For example, ['1960', '2023'] (full dataset), or ['1990', '2010'] (partial dataset).
12. Prompt the user to input countries one-at-a-time, where detailed analysis will be performed on that subset. E.g., ['Canada', 'Germany', 'Ireland', 'South Korea'].
13. Plot trends and aggregate statistics for the selected countries across the selected time frame.
14. Write the final MultiIndex data frame to an Excel file.

## 4 Project Requirements

This section describes how and where this project meets the stated requirements.

Project Requirement	Description of Meeting Requirement
Must use at least 3 separate dataset files.	Project uses 3 datasets from separate csv files, containing data on Total GDP, GDP Per Capita, and Gini Coefficient.
Final data must have at least 10 columns and 200 rows.	Final data frame shape is: (732,64)
Program must not modify excel files directly.	Program reads in and makes copies of original data, preserving the integrity of the data.
Must not hard-code / copy paste values.	All data is read in and modified via variables.

Must use at least 2 merge/joins on the data.	We do not use the merge or join functions explicitly. This is because the shape of each file is 2D, and it makes more sense to stack vertically. Therefore, we use the <i>pd.concat()</i> method to stack our data, and the MultiIndex of ['Metric', 'Country'].
You must create a hierarchical index of at least two levels (row or column).	We use a MultiIndex on the rows with 3 levels: ['Metric', 'Classification', and 'Country']
All data should be presented in the correctly sorted order, depending on the index.	The countries are sorted alphabetically, and this applies in each Metric.
You may not use global variables. You must import the data within your main function.	No global variables used. Data imported from csv files in <i>main()</i> .
Remember to check for null values or data mismatches.	Data is combined using inner join on both rows and columns, ensuring consistent shape of dataset. Additionally, rows that do not meet a specified threshold of valid data values are dropped. Remaining missing values are ignored.
<b>Your application must return useful information. Design an interface that allows users to search based on some sort of criteria or keywords.</b> The user must provide at least two pieces of information/selection (e.g. "school name" and "grade").	User interface asks user to specify a time range for detailed analysis by entering a Start Year: End Year: Then, the user is asked to enter an arbitrary number of countries one-at-a-time for analysis. For example, user could enter Country 1: Canada ['Canada'] Country 2: Germany ['Canada', 'Germany'] Country 3: Ireland ['Canada', 'Germany', 'Ireland']
Give the user clear input instructions. If an invalid entry is given, use try/except statements to handle the error and continue to prompt for user input.	User can press 1 for a list of valid country names to enter. Two functions: <i>check_years()</i> and <i>check_country()</i> use error handling to validate user input.
Any output information must be clearly defined using printed headers (DataFrame tables) or sentences (scalar values).	General information output is separated by header lines, user specific information is plotted in figures with appropriate title and legend.
Use the describe method to print aggregate stats for the entire dataset.	Our <i>describe_all()</i> function uses the <i>describe()</i> method applied to each Metric in the data frame.

Add at least two columns to the combined dataset.	We add a column 'Classification' as an additional layer of row indexing. And we also derive a new value (Gini Dollars) for each year/country. However, this is added as additional rows vertically. Again, this makes more sense given the shape of our data.
Use an aggregation computation for a subset of the data.	For the set of user selected countries, the mean values are plotted along with the individual countries in <i>plotter()</i> .
Use a masking operation.	In <i>df_index_inner_join()</i> masking is used to only include countries that are present in all datasets.
Use the 'groupby' operation at least once.	In the <i>describe_all()</i> function, groupby is used to access the metric subset of the data frame to use in the <i>describe()</i> method.
Create and print a pivot table.	A pivot table is created of averaged values for a specific year (default is end year of user input range), pivoting on 'Metric' and 'Classification'.
Include at least two user-defined functions or a class that contains two methods.	<i>drop_countries_by_nan()</i> takes in a threshold as a percentage, and drops a country if the percent of nan values exceeds the threshold. <i>plotter()</i> takes in the user defined lists of years and countries, and plots the various metrics for the specified data.
export your entire merged, hierarchical dataset to an Excel file in the working directory.	Final data frame is exported to excel as "final_output.xlsx". The file contains 3 sheets, with data summaries of all countries, data summaries of the user selected countries, and the full MultiIndex data, respectively.
Use your data to create at least one plot using Matplotlib. Save the plot as a .png file and upload it to the repository.	Program generates 4 plots, one for each metric showing the user specified groups of countries.
All classes, methods, and functions must contain docstring documentation	All functions contain docstrings, and steps in the main() are commented.

## 5 Code Output Screenshots

```
mwatson/VScode-projects/ENSF692/Project/leprechaun_economics/data_analysis.py
***** Cleaning Data *****
Original data size: (579, 64)
***** DROPPING COUNTRIES *****
Dropping Afghanistan
- 40 missing values in Total GDP.
- 40 missing values in GDP Per Capita.
Dropping Djibouti
- 53 missing values in Total GDP.
- 53 missing values in GDP Per Capita.
Dropping Eritrea
- 44 missing values in Total GDP.
- 44 missing values in GDP Per Capita.
Dropping Liechtenstein
- 63 missing values in Total GDP.
- 63 missing values in GDP Per Capita.
Dropping Montenegro
- 37 missing values in Total GDP.
- 37 missing values in GDP Per Capita.
Dropping Palestine
- 34 missing values in Total GDP.
- 34 missing values in GDP Per Capita.
Dropping San Marino
- 38 missing values in Total GDP.
- 38 missing values in GDP Per Capita.
Dropping Serbia
- 35 missing values in Total GDP.
- 35 missing values in GDP Per Capita.
Dropping South Sudan
- 56 missing values in Total GDP.
- 56 missing values in GDP Per Capita.
Dropping Yemen
- 35 missing values in Total GDP.
- 30 missing values in GDP Per Capita.
*****
New data size: (549, 64)
*****
```

```

*****
***** General Stats *****
['1960' '1970' '1980' '1990' '2000' '2010' '2020' '2023']
Describing GDP Per Capita metric by decade for: all countries
      1960      1970      1980      ...      2010      2020      2023
count  104.000000  135.000000  154.000000  ...  183.000000  183.000000  179.000000
mean   3339.701945  7334.488889  8617.032468  ...  11254.185792  12046.699454  14089.162011
std    7230.094887  14854.512267  18063.844397  ...  20714.231863  22165.731941  26083.573604
min      0.002300    13.000000    18.000000  ...   29.000000    25.000000    25.000000
25%     82.500000    96.000000   118.500000  ...  155.000000   195.500000   212.500000
50%    188.500000   242.000000   320.000000  ...   502.000000   557.000000   639.000000
75%    630.750000  11300.000000  11700.000000  ...  14700.000000  16700.000000  18950.000000
max   39900.000000  75800.000000  118000.000000  ...  142000.000000  161000.000000  225000.000000

[8 rows x 8 columns]
Describing Gini Coeff metric by decade for: all countries
      1960      1970      1980      1990      2000      2010      2020      2023
count  183.000000  183.000000  183.000000  183.000000  183.000000  183.000000  183.000000  183.000000
mean   41.408743  40.596175  39.329508  39.46776  39.968852  38.882514  38.646995  40.057923
std    10.346382  10.600157  10.227844  9.98107  8.636013  7.906516  7.842970  8.031369
min    19.000000  17.500000  19.300000  19.10000  23.800000  24.800000  25.200000  26.100000
25%    35.100000  33.850000  32.100000  32.50000  33.550000  33.150000  33.450000  34.800000
50%    40.700000  39.900000  38.100000  37.90000  38.400000  37.500000  37.200000  38.000000
75%    46.700000  46.600000  45.500000  45.30000  44.500000  43.700000  43.100000  44.750000
max    73.500000  70.800000  66.900000  68.90000  65.600000  63.400000  64.000000  66.200000
Describing Total GDP metric by decade for: all countries
      1960      1970      1980      ...      2010      2020      2023
count  1.040000e+02  1.350000e+02  1.540000e+02  ...  1.830000e+02  1.830000e+02  1.790000e+02
mean   9.700888e+10  1.245126e+11  1.589719e+11  ...  3.490473e+11  4.446725e+11  5.136243e+11
std    3.626224e+11  4.937976e+11  6.398188e+11  ...  1.413254e+12  1.880381e+12  2.163662e+12
min    1.250000e+02  1.380000e+07  1.170000e+07  ...  3.060000e+07  4.400000e+07  4.690000e+07
25%    2.472500e+09  2.110000e+09  2.450000e+09  ...  7.090000e+09  1.011500e+10  1.170000e+10
50%    7.660000e+09  8.580000e+09  1.090000e+10  ...  2.730000e+10  4.030000e+10  4.520000e+10
75%    4.965000e+10  5.055000e+10  6.725000e+10  ...  1.760000e+11  2.080000e+11  2.455000e+11
max    3.430000e+12  5.170000e+12  7.060000e+12  ...  1.630000e+13  1.970000e+13  2.210000e+13

```

```

Describing GDP Per Capita metric by decade for: all countries
      1960      1970      1980      ...      2010      2020      2023
count  104.000000  135.000000  154.000000  ...  183.000000  183.000000  179.000000
mean   3339.701945  7334.488889  8617.032468  ...  11254.185792  12046.699454  14089.162011
std    7230.094887  14854.512267  18063.844397  ...  20714.231863  22165.731941  26083.573604
min      0.002300   13.000000   18.000000  ...   29.000000   25.000000   25.000000
25%     82.500000   96.000000  118.500000  ...  155.000000  195.500000  212.500000
50%    188.500000  242.000000  320.000000  ...  502.000000  557.000000  639.000000
75%    630.750000 11300.000000 11700.000000  ... 14700.000000 16700.000000 18950.000000
max   39900.000000 75800.000000 118000.000000  ... 142000.000000 161000.000000 225000.000000

```

[8 rows x 8 columns]

```

Describing Gini Coeff metric by decade for: all countries
      1960      1970      1980      1990      2000      2010      2020      2023
count  183.000000  183.000000  183.000000  183.000000  183.000000  183.000000  183.000000  183.000000
mean    41.408743  40.596175  39.329508  39.46776  39.968852  38.882514  38.646995  40.057923
std     10.346382  10.600157  10.227844  9.98107  8.636013  7.906516  7.842970  8.031369
min     19.000000  17.500000  19.300000  19.10000  23.800000  24.800000  25.200000  26.100000
25%     35.100000  33.850000  32.100000  32.50000  33.550000  33.150000  33.450000  34.800000
50%     40.700000  39.900000  38.100000  37.90000  38.400000  37.500000  37.200000  38.000000
75%     46.700000  46.600000  45.500000  45.30000  44.500000  43.700000  43.100000  44.750000
max     73.500000  70.800000  66.900000  68.90000  65.600000  63.400000  64.000000  66.200000

```

```

Describing Gini Dollars metric by decade for: all countries

```

```

      1960      1970      1980      ...      2010      2020      2023
count  104.000000  135.000000  154.000000  ...  183.000000  183.000000  179.000000
mean   2131.586061  4776.383474  5737.938786  ...  7412.884257  7909.749984  9061.642989
std    4706.221057  9786.856757  12116.999123  ... 13928.888595  14706.946944  17095.170030
min      0.001334   8.099000   11.142000  ...   18.473000   14.950000   14.125000
25%     44.215500   45.934000   66.486000  ...   95.631000  114.214500  122.200500
50%     92.988500  130.284000  158.523500  ...  273.884000  308.649000  350.132000
75%    384.008250  6718.100000  7442.625000  ... 10037.300000 10705.750000 11651.700000
max   24777.900000 50634.400000 80594.000000  ... 92300.000000 104972.000000 146700.000000

```

[8 rows x 8 columns]

```

Describing Total GDP metric by decade for: all countries
      1960      1970      1980      ...      2010      2020      2023
count  1.040000e+02 1.350000e+02 1.540000e+02  ...  1.830000e+02 1.830000e+02 1.790000e+02
mean   9.700888e+10 1.245126e+11 1.589719e+11  ...  3.490473e+11 4.446725e+11 5.136243e+11
std    3.626224e+11 4.937976e+11 6.398188e+11  ...  1.413254e+12 1.880381e+12 2.163662e+12
min    1.250000e+02 1.380000e+07 1.170000e+07  ...  3.060000e+07 4.400000e+07 4.690000e+07
25%    2.472500e+09 2.110000e+09 2.450000e+09  ...  7.090000e+09 1.011500e+10 1.170000e+10
50%    7.660000e+09 8.580000e+09 1.090000e+10  ...  2.730000e+10 4.030000e+10 4.520000e+10
75%    4.965000e+10 5.055000e+10 6.725000e+10  ...  1.760000e+11 2.080000e+11 2.455000e+11
max    3.430000e+12 5.170000e+12 7.060000e+12  ...  1.630000e+13 1.970000e+13 2.210000e+13

```

[8 rows x 8 columns]



```
Enter a time frame to look at. Valid range is: 1960 to 2023
Start year: 1959
End year: 2023
'Invalid year range. Years must be in proper range, and start year must come before end year.'
Enter a time frame to look at. Valid range is: 1960 to 2023
Start year: 1965
End year: 2023
Enter a country to include in the analysis. Multiple countries may be added.
Enter 1 for a list of valid countries to enter
Enter 0 to analyze current list of countries.
Enter country number 1 : Canada
Current country list: ['Canada']
Enter a country to include in the analysis. Multiple countries may be added.
Enter 1 for a list of valid countries to enter
Enter 0 to analyze current list of countries.
Enter country number 2 : US
'Country not in list. Try again.'
Current country list: ['Canada']
Enter a country to include in the analysis. Multiple countries may be added.
Enter 1 for a list of valid countries to enter
Enter 0 to analyze current list of countries.
Enter country number 2 : USA
Current country list: ['Canada', 'USA']
Enter a country to include in the analysis. Multiple countries may be added.
Enter 1 for a list of valid countries to enter
Enter 0 to analyze current list of countries.
Enter country number 3 : Ireland
Current country list: ['Canada', 'USA', 'Ireland']
Enter a country to include in the analysis. Multiple countries may be added.
Enter 1 for a list of valid countries to enter
Enter 0 to analyze current list of countries.
Enter country number 4 : Germany
Current country list: ['Canada', 'USA', 'Ireland', 'Germany']
Enter a country to include in the analysis. Multiple countries may be added.
Enter 1 for a list of valid countries to enter
Enter 0 to analyze current list of countries.
Enter country number 5 : 0
(['Canada', 'USA', 'Ireland', 'Germany'], ['1965', '2023'])
```

```

Describing GDP Per Capita metric by decade for: ['Canada', 'USA', 'Ireland', 'Germany']
      1965      1970      1980      1990      2000      2010      2020      2023
count    4.000000    4.000000    4.000000    4.000000    4.000000    4.000000    4.000000    4.000000
mean   16600.000000  19250.000000  24900.000000  31025.000000  41050.000000  44875.000000  56525.000000  61575.000000
std     5260.544965   5466.56504    6070.145523   6472.184072   5908.468499   6363.109827   18678.931982  22436.484425
min    10100.000000  12300.000000  16900.000000  23500.000000  35100.000000  38500.000000  42400.000000  44300.000000
25%    13925.000000  16650.000000  22225.000000  28150.000000  37200.000000  40525.000000  42400.000000  44450.000000
50%    16900.000000  19750.000000  25800.000000  30700.000000  40250.000000  44100.000000  50950.000000  55200.000000
75%    19575.000000  22350.000000  28475.000000  33575.000000  44100.000000  48450.000000  65075.000000  72325.000000
max     22500.000000  25200.000000  31100.000000  39200.000000  48600.000000  52800.000000  81800.000000  91600.000000

Describing Gini Coeff metric by decade for: ['Canada', 'USA', 'Ireland', 'Germany']
      1965      1970      1980      1990      2000      2010      2020      2023
count    4.000000    4.000000    4.000000    4.000000    4.000000    4.000000    4.000000    4.000000
mean     35.300000    34.400000    33.175000    33.575000    33.800000    34.050000    35.300000    37.575000
std      3.332667     3.165438     3.271468     4.096645     4.675468     4.192454     4.938286     5.635823
min      32.200000    29.700000    28.900000    29.200000    28.800000    30.300000    32.000000    33.600000
25%      32.950000    34.125000    31.600000    30.625000    31.950000    31.800000    32.450000    34.500000
50%      34.700000    35.650000    33.700000    33.600000    33.150000    32.950000    33.300000    35.400000
75%      37.050000    35.925000    35.275000    36.550000    35.000000    35.200000    36.150000    38.475000
max      39.600000    36.600000    36.400000    37.900000    40.100000    40.000000    42.600000    45.900000

Describing Gini Dollars metric by decade for: ['Canada', 'USA', 'Ireland', 'Germany']
      1965      1970      1980      1990      2000      2010      2020      2023
count    4.000000    4.000000    4.000000    4.000000    4.000000    4.000000    4.000000    4.000000
mean   10796.450000  12607.400000  16672.125000  20557.150000  26980.975000  29422.575000  36584.650000  38459.475000
std     3540.650653   3470.162814   4157.349144   3954.041875   2147.123043   2695.946276  12993.508639  15267.287417
min     6100.400000   7908.900000  10748.400000  15016.500000  24991.200000  26834.500000  27984.000000  28480.000000
25%     9254.300000  11520.450000  15485.100000  19524.825000  25207.275000  27226.225000  28429.200000  28782.700000
50%    11365.200000  13145.950000  17847.000000  21434.450000  26910.650000  29518.400000  31365.300000  32267.750000
75%    12907.350000  14232.900000  19034.025000  22466.775000  28684.350000  31714.750000  39520.750000  41944.525000
max    14355.000000  16228.800000  20246.100000  24343.200000  29111.400000  31819.000000  55624.000000  60822.400000

Describing Total GDP metric by decade for: ['Canada', 'USA', 'Ireland', 'Germany']
      1965      1970      1980      1990      2000      2010      2020      2023
count    4.000000e+00  4.000000e+00  4.000000e+00  4.000000e+00  4.000000e+00  4.000000e+00  4.000000e+00  4.000000e+00
mean     1.481250e+12  1.768350e+12  2.418700e+12  3.277625e+12  4.475500e+12  5.266250e+12  6.309500e+12  7.014000e+12
std      1.988743e+12  2.339439e+12  3.185343e+12  4.442975e+12  6.251289e+12  7.453974e+12  9.018496e+12  1.014308e+13
min      2.900000e+10  3.640000e+10  5.780000e+10  8.250000e+10  1.620000e+11  2.150000e+11  4.080000e+11  4.860000e+11
25%      2.817500e+11  3.518500e+11  5.222000e+11  6.791250e+11  9.105000e+11  1.103750e+12  1.309500e+12  1.456500e+12
50%      7.580000e+11  9.335000e+11  1.278500e+12  1.619000e+12  2.020000e+12  2.275000e+12  2.565000e+12  2.735000e+12
75%      1.957500e+12  2.350000e+12  3.175000e+12  4.217500e+12  5.585000e+12  6.437500e+12  7.565000e+12  8.292500e+12
max      4.380000e+12  5.170000e+12  7.060000e+12  9.790000e+12  1.370000e+13  1.630000e+13  1.970000e+13  2.210000e+13

```

```

Pivot Tabe of 2023 data, pivoting on Classification and Metric
Classification Middleweight Heavyweight Superpower
Metric
GDP Per Capita 9.160000e+04 4.440000e+04 6.590000e+04
Gini Coeff 3.360000e+01 3.540000e+01 4.590000e+01
Gini Dollars 6.082240e+04 2.868180e+04 3.565190e+04
Total GDP 4.860000e+11 2.735000e+12 2.210000e+13

```



