THE USER GUIDE TO
An Visual Exploration of World Economy and Demographics

YICHAO CHEN, YC1228
ZOE MA, YM910
LIWEI SONG, LS4408

NYU Data Science Institute
December 16, 2016

Abstract

World economy and demographics is a complex and confounding topic to explore, and by simply looking at data it is hard to obtain any valuable information from it. Hence, we think it is meaningful to develop an understandable visualization tool to help the process of exploration for people, especially for those without previous data analysis experience. This program takes a tuple including a list of countries from 208 total, a list of years ranging from 1991 to 2016, a list of features from 87 total, and a plot type as input, and gives interactive plots accordingly for the tuple of interests. Our data contains Global Economics Monitor [1] and Gender Statistics [2] from World Bank Open Data Catalog.

[1] `http://databank.worldbank.org/data/databases.aspx?qterm=economic&pagenumber=1`
[2] `http://databank.worldbank.org/data/reports.aspx?source=gender-statistics#`

Contents

1 Prerequisite Installation

The program requires following installations and setup before it can be successfully run. For best visualization results, please run it in iPython notebook.

- The program requires python 3.5 [3] or above.
- We recommend to use iPython Notebook to run the program. To install iPython and Jupyter Notebook, and Anaconda [4]. Please enter following command in your terminal on Mac or Command Prompt on Windows:
  `pip install ipython`
  and when you already install Anaconda, use following command to install Jupyter Notebook:
  `jupyter notebook`
- Following packages are required:
  pandas, numpy, matplotlib, plotly, pickle [5]
  To install packages, please enter following command in your terminal on Mac or Command Prompt on Windows:
  `pip install pandas, numpy, matplotlib, plotly`
- Download files:
  The whole package includes one start ipynb file, 3 data frame files, 9 python files, a `data` folder of 87 Excel files, an Excel file for country name and country code spreadsheet, and a pdf file: user_guide.pdf. Please put all files in one folder in order to run the program properly.
  - python files:
    - `Run_test_project.ipynb`: open this file to start the program.
    - `class_function.py`: this file contains all functions used in this program, including functions to justify errors, build up the full database, define the format of country input, year input, and feature input, etc.
    - `country_input.py`: this file gives instructions on how to input country codes, takes a list of country code as input and returns a list of string defined in `class_function.py`.
    - `year_input.py`: this file gives instructions on how to input year, takes an interval as input and returns an integer interval defined in `class_function.py`.

[3] download from here: `https://www.python.org/downloads/`
[4] download from here: `https://www.continuum.io/downloads`
[5] pickle is pre-installed with python.

    * `feature_input.py`: this file gives instructions on how to input features, takes a list of features as input and returns a list of string defined in `class_function.py`.
    * `db_input.py`: this file saves all data into the database and secures it from any user modification; also it can be directly used in SQL or SQLite in python for future use.
    * `plot_input.py`: this file gives instructions on how to input types of plot, takes an integer from 1 to 9 as input and returns the corresponding plot.
    * `plot.py`: this file takes selected data and returns interactive plots that can move around across range of years and show data of a specific data points.
    * `1007_p_main.py`: this file is the main program.
    * `test.py`: this file is the unit test for the program.
- self-defined dataframe files:
  We use *pickle* to save the country codes to a dataframe, and further organize it to be a dictionary with key to be the initials for better search results when you input the list of countries. We also save the list of years as a dataframe.
    * `contry_dict.df`: this file is the saved dictionary with key to be the initials.
    * `contryname_list.df`: this file is the saved dataframe of country codes.
    * `year_list.df`: this file is the saved dataframe of list of years from 1991 to 2016.
- data files:
  All data are located in the `data` folder. It contains 87 Excel files, each one represents one feature and is named precisely what the feature is. The column index are in form of country code and the row index are in form of year. Example:
    * `total_reserve.csv`: this economics data file records Total Reserves for 208 countries from 1991 to 2018.
    * `core_cpi_n.csv`: this economics data file records the Core Consumer Price Index (CPI excludes food and energy) to measure core inflation for 208 countries from 1991 to 2018.
    * `labor_force_participation_rate.csv`: this demographic data file records the participation rate that measures the active portion of an economy's labor force for 208 countries from 1991 to 2018.
- database:
    * `worldbank.db`: this file will occur when you follow the program instructions and install the database.
- country code list:
    * `country_name_code`: this file is a list of country names and corresponding country code for you to check, for reference.

2 Launching the Program

To begin the program, please open the iPython Notebook `Run_test_project.ipynb` in the same directory where you store the downloaded files, and run the following command in the

first cell:

```
% run 1007_p_main
```

Make sure you are working in the directory that you downloaded.
We give an example to run through the program.

1. After you run the first command, the main menu will appear:
   ```
   Welcome to the world bank data explorer
   countries,year,features
   This is the variables input menu, and you have 5 options:
   option1:enter database builder, option2:  choose interested
   countries option3:  choose interested features
   option4:  choose interested year(s).Please choose values
   for countries, features and years to next stage
   Or you could enter Q to exit the system
   enter C to continue enter your options from [1,2,3,4,C,Q]:
   ```
   If this is your first time to launch this program, please choose option 1 to install the database. You can modify your input or quit anytime.
2. Now you can choose from option 2, 3, 4 to input a list of countries, a list of features, or an year interval. We follow the order of option 2, 3, 4 and choose option 2 here as an illustration of how the program works. After you input 2, following message will show up:
   ```
   Enter countries names in forms of list,such as [A, B, C]
   No white space between country names
   Please enter your interested countries!  Enter Help for country name
   Enter back to return to main menu:
   ```
   You can enter `A` if you want to choose all 208 countries. If you are not sure what country to choose, or not sure how to input country code, please enter `H` for instructions, and the program will return following message:
   ```
   Please enter the initial your country, and enter back to go back:
   ```
   Now if we input `A`, the program will list all country codes start with A:
   ```
   ['AFG', 'ALB', 'AGO', 'ATG', 'ARG', 'ARM', 'ABW', 'AUS', 'AUT', 'AZE', 'ARE']
   Please enter the initial your country, and enter back to go back:
   ```
   You can always enter `B` to return to the main menu.
3. We input three random country codes from the list given above, say `[AFG,ALB,AGO]`. Note that your input should strictly follow the format: begin with square bracket, enter your country codes without any space, then end with square bracket. Following message will appear:
   ```
   Do you want to edit your input?Y to edit,Q to exit the system,Anything else
   back to main menu
   ```
   You can always choose `Y` to edit your inputs for country, year or feature anytime. Press any button to return to the main menu, and we will go to option 3. If you choose `continue` now, it will reminds you that:
   ```
   Could not proceed to the next stage, as one of the three attributes might be
   empty
   ```

4. When you choose option 3, the program will show all available features. Please strictly follow the format to begin your input with square bracket, enter your choice of features from above without any space, then end with square bracket. Please choose at most 5 features to keep the plot interpretable. Say we input `[prevalence_of_underweight_f,total_reserve]` here, and return to the main menu again.
5. Now we continue with option 4. It will show following messages:
   ```
   Enter your interested years, and enter back to go back
   Enter your interested years, such as [1991], [1991,1992].:
   Available years are from 1991 to 2016, and enter Q to quit
   ```
   Please input an interval, following the format mentioned above. Say we input `[1991,2016]`.
6. Now we have all inputs ready, and the program will print your choices for review. Here we have:
   ```
   Here are your interested countries list
   ['AFG', 'ALB', 'AGO']
   Here are your interested features list
   ['prevalence_of_underweight_f', 'total_reserve']
   Here are your interested years list
   [1991, 2016]
   ```
   You can still enter `R` to return to the main menu and edit your inputs.
7. Now you can choose a plot type from 9 available types: `[1.time_series_plot, 2.barplot, 3.scatter_plot, 4.histogram, 5.boxplot_year, 6.boxplot_country, 7.heatmap, 8.choropleth, 9.pie_chart]`. Enter the number of your choice of plot type, the plots will present accordingly. You can quit the program from here, return to choose another type, or press `R` twice to return to the main menu and modify your choice of feature, year and country.

3 Figure Illustration

We show all 9 types of plots as an illustration for feature `total_reserve` for countries `['CHN', 'USA', 'GBR']` and years `[1991, 2016]`. Note that if you see large space, it means that the data is lacked in the original dataset. To ensure the completeness of your examination, we decide not to replace the missing data with 0.

- Time-Series Plot [Figure.1]
- Bar Plot [Figure.2]
- Scatter Plot [Figure.3]
- Histogram [Figure.4]
- Box Plot in Terms of Year [Figure.5]
- Box Plot in Terms of Country [Figure.6]
- Heatmap [Figure.7]
- Choropleth in year 2000 [Figure.8]

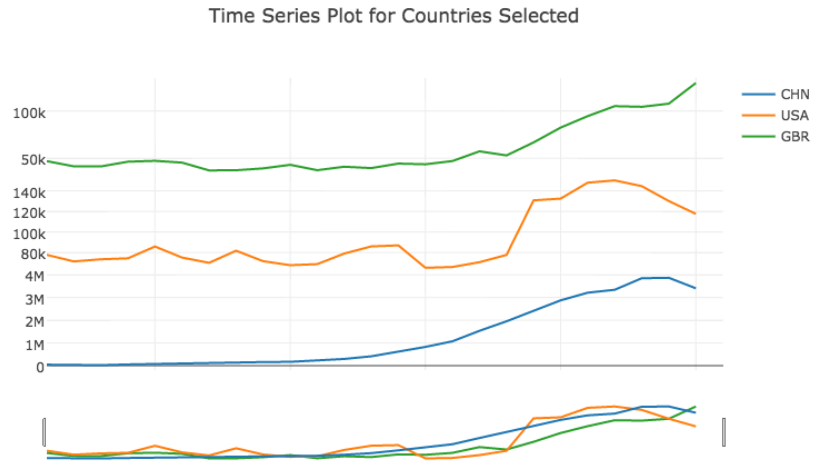

Figure 1: An example of time-series from 1991 to 2016 for countries China, UK, and USA. It is an interactive plot in ipython Notebook, you can slide on the bottom year range to see how the plot is changing over time.

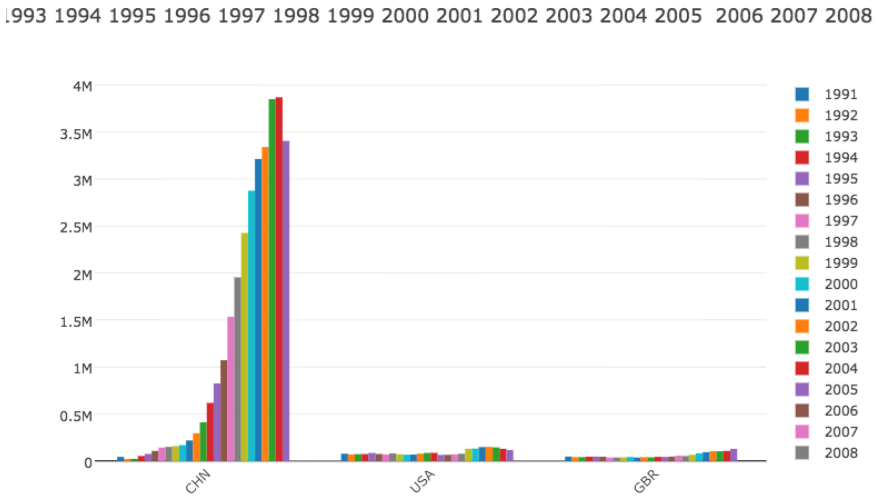

Figure 2: An example of bar plot from 1991 to 2016 for countries China, UK, and USA.

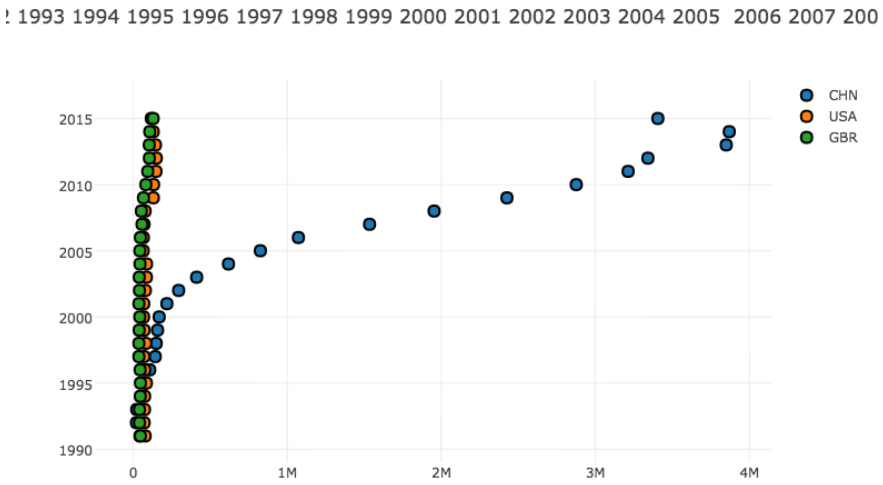

Figure 3: An example of scatter plot from 1991 to 2016 for countries China, UK, and USA.

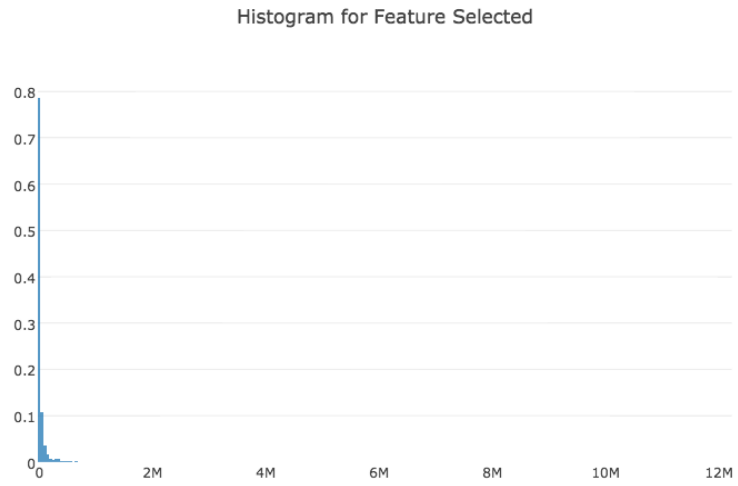

Figure 4: An example of histogram from 1991 to 2016 for countries China, UK, and USA. It shows the density of the selected feature in range.

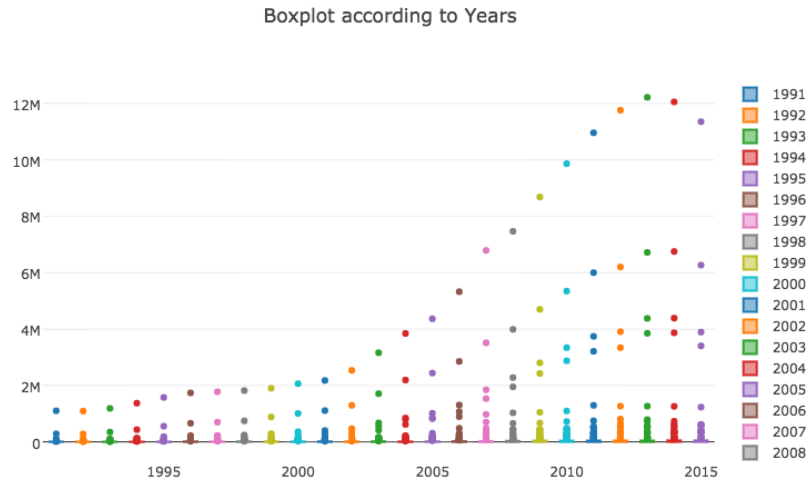

Figure 5: An example of box plot in terms of years for the countries you selected in one year. If you input a year interval, it will return a plot for every year.

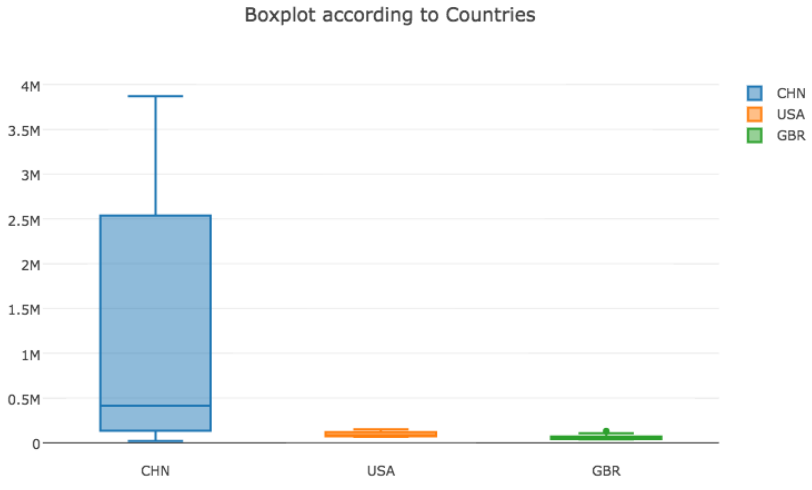

Figure 6: An example of box plot in terms of country from 1991 to 2016 for one country. If you input a list of countries, it will return a plot for every country.

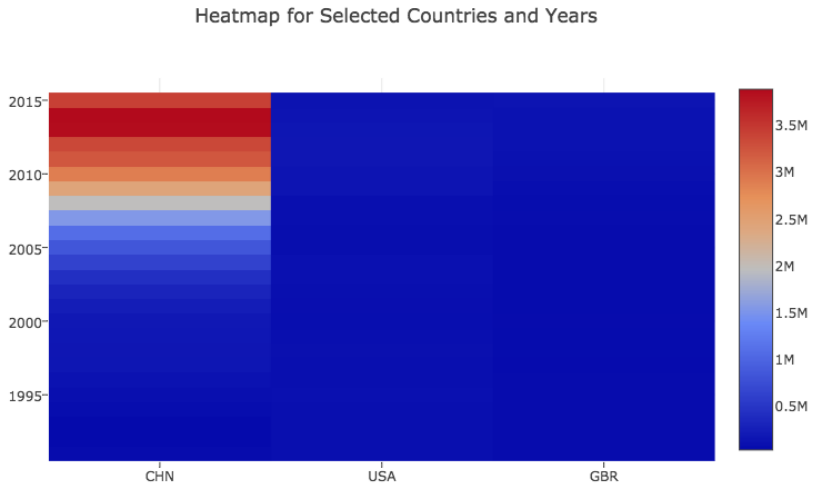

Figure 7: An example of heatmap from 1991 to 2016 for countries China, UK, and USA. The plot summarize the relationships of distributions to the selected features.

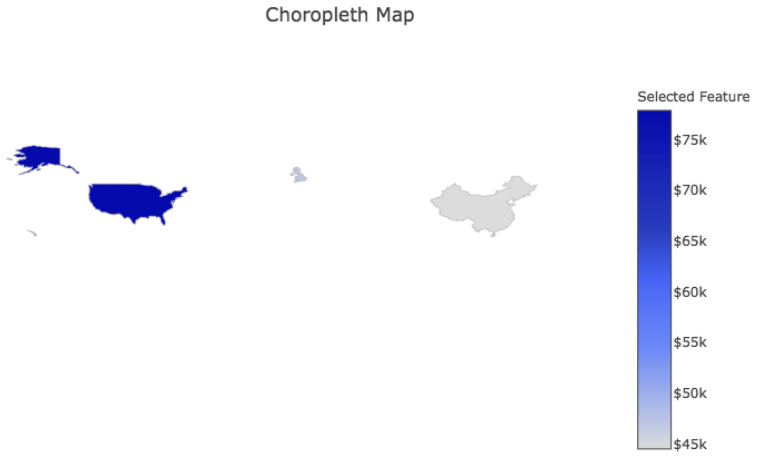

Figure 8: An example of chronopleth of total reserve. It can plot one year data, so if you input a year interval, it will return plots of each year.

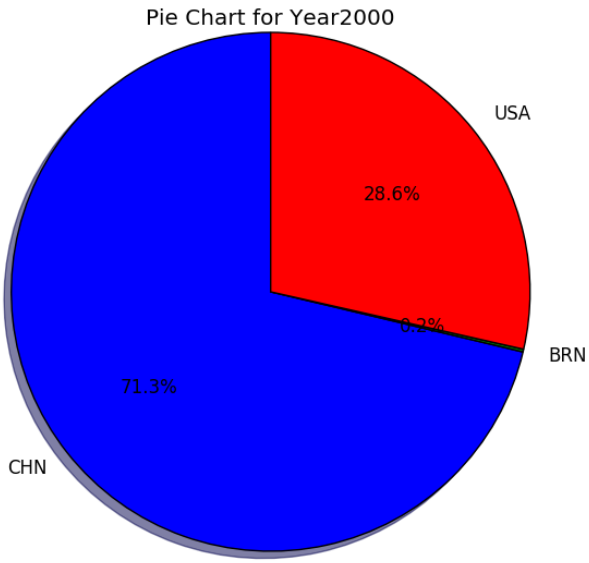

Figure 9: An example of pie chart of total reserve at year 2000. It can plot one year data, so if you input a year interval, it will return plots of each year.

- Pie Chart in year 2000 [Figure.9]