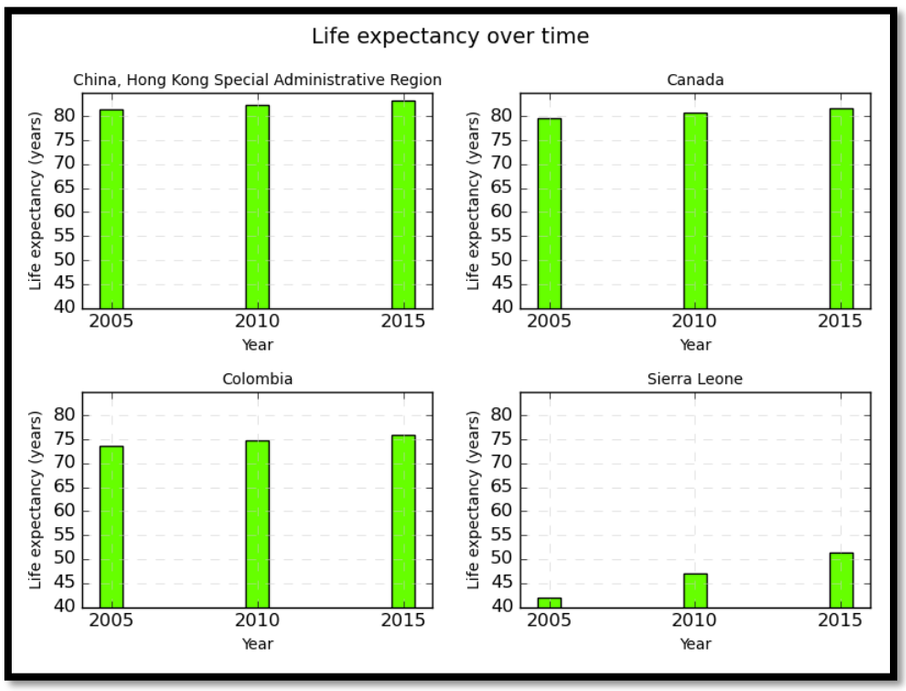
ENSF 592 L01 - Spring 2022

Programming Fundamentals for Data Engineers

Final Project

Kernel-Panic – Group 48



Kevin David Amado Rueda

kevin.amadorueda@ucalgary.ca

UCID 30169590

Tyson Trail

tyson.trail@ucalgary.ca

UCID 30146430

Source Code

https://github.com/MaanKhedr-teaching/final-project-kernel-panic

Chosen Dataset

We used two of the spreadsheets provided with the original repository with name:

- **un-population-dataset-1.xlsx -** Per each country and for some years, it contains the life expectancy, fertility rates and annual rates of increase in population.
- **un-population-dataset-2.xlsx -** Per each country and for some years, it contains the urban and capital city population.

And incorporated a third one:

- **un-m49.xlsx** - Per each country, it contains a unique identifier and associated data like regions and ISO codes.

  Citation:

  Standard country or area codes for statistical use (M49), Statistics Division from the Department of Economic and Social Affairs of the United Nations, June 2022. [Online]. Available: https://unstats.un.org/unsd/methodology/m49/overview/

The reason for using this third spreadsheet is to have an identifier that can be reliably used as index (primary key) in a Pandas “merge” operation. The original third spreadsheet included in the repository had country names mismatches, and thus the “merge” operation was unreliable.

The merged dataset looks like this:

Country or Area	Year	Global Name	Region Name	Sub-region Name	M49 Code	ISO-alpha2 Code	ISO-alpha3 Code	Capital city population (as a percentage of total population)	Capital city population (a
Afghanistan	2005	World	Asia	Southern Asia	4	AF	AFG	11.6	
	2010	World	Asia	Southern Asia	4	AF	AFG	11.4	
	2015	World	Asia	Southern Asia	4	AF	AFG	11	
Albania	2005	World	Europe	Southern Europe	8	AL	ALB	12.1	
	2010	World	Europe	Southern Europe	8	AL	ALB	13.9	
	2015	World	Europe	Southern Europe	8	AL	ALB	15.4	
Algeria	2005	World	Africa	Northern Africa	12	DZ	DZA	6.9	
	2010	World	Africa	Northern Africa	12	DZ	DZA	6.7	
	2015	World	Africa	Northern Africa	12	DZ	DZA	6.5	
Angola	2005	World	Africa	Sub-Saharan Africa	24	AO	AGO	19.8	
	2010	World	Africa	Sub-Saharan Africa	24	AO	AGO	22.7	
	2015	World	Africa	Sub-Saharan Africa	24	AO	AGO	25.2	
Argentina	2005	World	Americas	Latin America and the Caribbean	32	AR	ARG	34.1	
	2010	World	Americas	Latin America and the Caribbean	32	AR	ARG	34.6	
	2015	World	Americas	Latin America and the Caribbean	32	AR	ARG	33.9	
Armenia	2005	World	Asia	Western Asia	51	AM	ARM	36.5	
	2010	World	Asia	Western Asia	51	AM	ARM	37	
	2015	World	Asia	Western Asia	51	AM	ARM	36.7	
Australia	2005	World	Oceania	Australia and New Zealand	36	AU	AUS	1.8	
	2010	World	Oceania	Australia and New Zealand	36	AU	AUS	1.8	
	2015	World	Oceania	Australia and New Zealand	36	AU	AUS	1.8	
Austria	2005	World	Europe	Western Europe	40	AT	AUT	19.9	
	2010	World	Europe	Western Europe	40	AT	AUT	20.6	
	2015	World	Europe	Western Europe	40	AT	AUT	21.1	
Azerbaijan	2005	World	Asia	Western Asia	31	AZ	AZE	21.9	
	2010	World	Asia	Western Asia	31	AZ	AZE	22.8	
	2015	World	Asia	Western Asia	31	AZ	AZE	22.9	
Bahrain	2005	World	Asia	Western Asia	48	BH	BHR	22.3	
	2010	World	Asia	Western Asia	48	BH	BHR	23.5	

Country and Year compose the index.

Task distribution

We pair programmed during most of the project:

- Kevin and Tyson worked individually on a proof of concept of merging the three datasets.
- Kevin was the driver and Tyson the observer while merging the three datasets.
- Tyson was the driver and Kevin the observer while working on the user input and part of the data analysis.
- Tyson worked alone on the remaining part of the data analysis.
- Kevin worked alone on the matplotlib, screenshots and report parts of the assignment.
- Tyson reviewed Kevin's pull requests and merged some of them, the other ones were self-merged by Kevin, but Tyson later reviewed them to synchronize with the project changes.
- Kevin reviewed and merged the pull requests from Tyson.

In general, both of us worked the same number of hours, and feel it was a fair distribution.

Management

We were very proactive so there was little need for management. We used two branches with names "kamadorueda" (for Kevin) and "tysontrail" (for Tyson). Then each of us created pull requests targeting the "main" branch with the changes made.

With this mechanism we avoided git conflicts and introduced the opportunity for the other person to review the changes and know what the other person was doing asynchronously. Like in real life.

User Interface

The program allows the user to query data from one of the available countries:

```
---
Available countries:

    ['Afghanistan' 'Albania' 'Algeria' 'Angola' 'Argentina' 'Armenia'
     'Australia' 'Austria' 'Azerbaijan' 'Bahrain' 'Bangladesh' 'Belarus'
     'Belgium' 'Bosnia and Herzegovina' 'Brazil' 'Bulgaria' 'Burkina Faso'
     'Burundi' 'Cambodia' 'Cameroon' 'Canada' 'Central African Republic'
     'Chad' 'Chile' 'China' 'China, Hong Kong Special Administrative Region'
     'China, Macao Special Administrative Region' 'Colombia' 'Congo'
     'Costa Rica' 'Croatia' 'Cuba' 'Czechia'
     "Democratic People's Republic of Korea"
     'Democratic Republic of the Congo' 'Denmark' 'Djibouti'
     'Dominican Republic' 'Ecuador' 'Egypt' 'El Salvador' 'Eritrea' 'Estonia'
     'Ethiopia' 'Finland' 'France' 'Gabon' 'Gambia' 'Georgia' 'Germany'
     'Ghana' 'Greece' 'Guatemala' 'Guinea' 'Guinea-Bissau' 'Haiti' 'Honduras'
     'Hungary' 'India' 'Indonesia' 'Iran (Islamic Republic of)' 'Iraq'
     'Ireland' 'Israel' 'Italy' 'Jamaica' 'Japan' 'Jordan' 'Kazakhstan'
     'Kenya' 'Kuwait' 'Kyrgyzstan' "Lao People's Democratic Republic" 'Latvia'
     'Lebanon' 'Liberia' 'Libya' 'Lithuania' 'Madagascar' 'Malawi' 'Malaysia'
     'Mali' 'Mauritania' 'Mexico' 'Mongolia' 'Morocco' 'Mozambique' 'Myanmar'
     'Nepal' 'Netherlands' 'New Zealand' 'Nicaragua' 'Niger' 'Nigeria'
     'North Macedonia' 'Norway' 'Oman' 'Pakistan' 'Panama' 'Papua New Guinea'
     'Paraguay' 'Peru' 'Philippines' 'Poland' 'Portugal' 'Puerto Rico' 'Qatar'
     'Republic of Korea' 'Republic of Moldova' 'Romania' 'Russian Federation'
     'Rwanda' 'Saudi Arabia' 'Senegal' 'Serbia' 'Sierra Leone' 'Singapore'
     'Slovakia' 'Somalia' 'South Africa' 'South Sudan' 'Spain' 'Sri Lanka'
     'Sudan' 'Sweden' 'Switzerland' 'Syrian Arab Republic' 'Tajikistan'
     'Thailand' 'Togo' 'Trinidad and Tobago' 'Tunisia' 'Turkmenistan'
     'Türkiye' 'Uganda' 'Ukraine' 'United Arab Emirates'
     'United Kingdom of Great Britain and Northern Ireland'
     'United States of America' 'Uruguay' 'Uzbekistan'
     'Venezuela (Bolivarian Republic of)' 'Viet Nam' 'Yemen' 'Zambia'
     'Zimbabwe']

Please enter a Country or Area: Canada
```

And then to query the data for one of the available years for that country:

```
---
Available Years:

    2005, 2010, 2015

Please choose one of the years above in order to display more stats: 2010
```

In general, the program acts proactively and shows the users which are the available countries and years, in order to avoid guessing and provide a better user experience.

Once the two parameters are introduced, the data available for the selected parameters is displayed:

```
---
Data available for Canada, year 2010:

    Global Name                                                          World
    Region Name                                                       Americas
    Sub-region Name                                          Northern America
    M49 Code                                                              124
    ISO-alpha2 Code                                                        CA
    ISO-alpha3 Code                                                       CAN
    Capital city population (as a percentage of total population)         3.6
    Capital city population (as a percentage of total urban population)   4.4
    Capital city population (thousands)                                1218.0
    Life expectancy at birth for both sexes (years)                    80.761
    Life expectancy at birth for females (years)                        82.99
    Life expectancy at birth for males (years)                          78.44
    Population annual rate of increase (percent)                        1.197
    Total fertility rate (children per women)                           1.636
    Urban population (percent)                                           80.9
    Life expectancy difference (years) from mean                    11.507365
    Total fertility difference (children per woman) from mean       -1.369499
    Name: 2010, dtype: object
```

```
---
Aggregate mean for years: 2005, 2010, 2015, for Canada:

    M49 Code                                                          124.000000
    Capital city population (as a percentage of total population)       3.566667
    Capital city population (as a percentage of total urban population)   4.400000
    Capital city population (thousands)                            1216.000000
    Life expectancy at birth for both sexes (years)                  80.747000
    Life expectancy at birth for females (years)                     82.980000
    Life expectancy at birth for males (years)                       78.423333
    Population annual rate of increase (percent)                      1.091000
    Total fertility rate (children per women)                         1.584333
    Urban population (percent)                                       80.766667
    Life expectancy difference (years) from mean                     11.493365
    Total fertility difference (children per woman) from mean        -1.421166
    dtype: float64
```

And then general statistics for all countries and years:

```
---
Aggregate stats for all years and all countries:

                                                count          mean          std  ...          50%          75%          max
    M49 Code                                    438.0  4.396164e+02   252.221769  ...   425.000000   646.000000   894.000000
    Capital city population (as a percentage of tot...  438.0  1.841301e+01    17.135869  ...    14.100000    22.525000   100.000000
    Capital city population (as a percentage of tot...  438.0  3.123562e+01    19.779358  ...    28.850000    41.800000   100.000000
    Capital city population (thousands)         438.0  3.236847e+03  4815.867473  ...  1605.500000  2938.750000  37256.000000
    Life expectancy at birth for both sexes (years)     438.0  6.925363e+01     9.579859  ...    72.098000    76.188750     83.404000
    Life expectancy at birth for females (years)        438.0  7.177032e+01    10.086036  ...    75.140000    79.082500     86.470000
    Life expectancy at birth for males (years)          438.0  6.676911e+01     9.192660  ...    68.885000    73.767500     80.500000
    Population annual rate of increase (percent)        438.0  1.543753e+00     1.660515  ...     1.351000     2.479500     15.263000
    Total fertility rate (children per women)           438.0  3.005499e+00     1.651245  ...     2.445000     4.177250      7.650000
    Urban population (percent)                  438.0  5.851781e+01    23.054081  ...    59.400000    77.200000     100.000000
    Life expectancy difference (years) from mean        438.0  2.092694e-14     9.579859  ...     2.844365     6.935115     14.150365
    Total fertility difference (children per woman)...  438.0  8.111218e-17     1.651245  ...    -0.560499     1.171751      4.644501

    [12 rows x 8 columns]
```

Pandas describe method

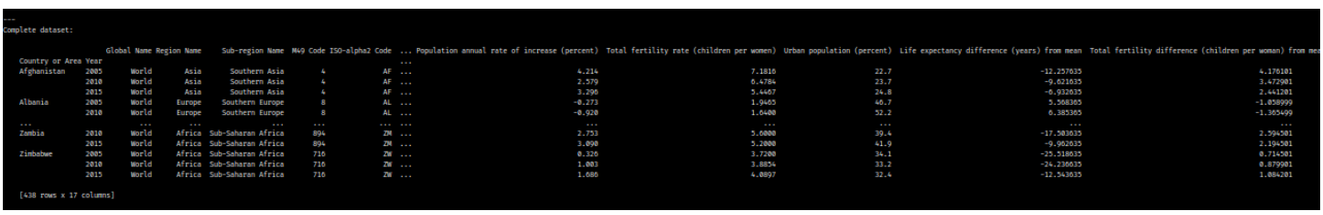
```
---
Fastest growing countries by average annual rate of increase (percent):

    Country or Area
    Qatar                     9.771333
    United Arab Emirates      7.224667
    Bahrain                   4.831667
    Oman                      4.214000
    Kuwait                    4.192333
                                ...
    Bosnia and Herzegovina   -0.598333
    Bulgaria                 -0.700667
    Romania                  -0.701667
    Latvia                   -1.179333
    Lithuania                -1.184333
    Name: Population annual rate of increase (percent), Length: 146, dtype: float64
```

“Group By” and Aggregation (maximum value) of the “average annual rate of increase” column

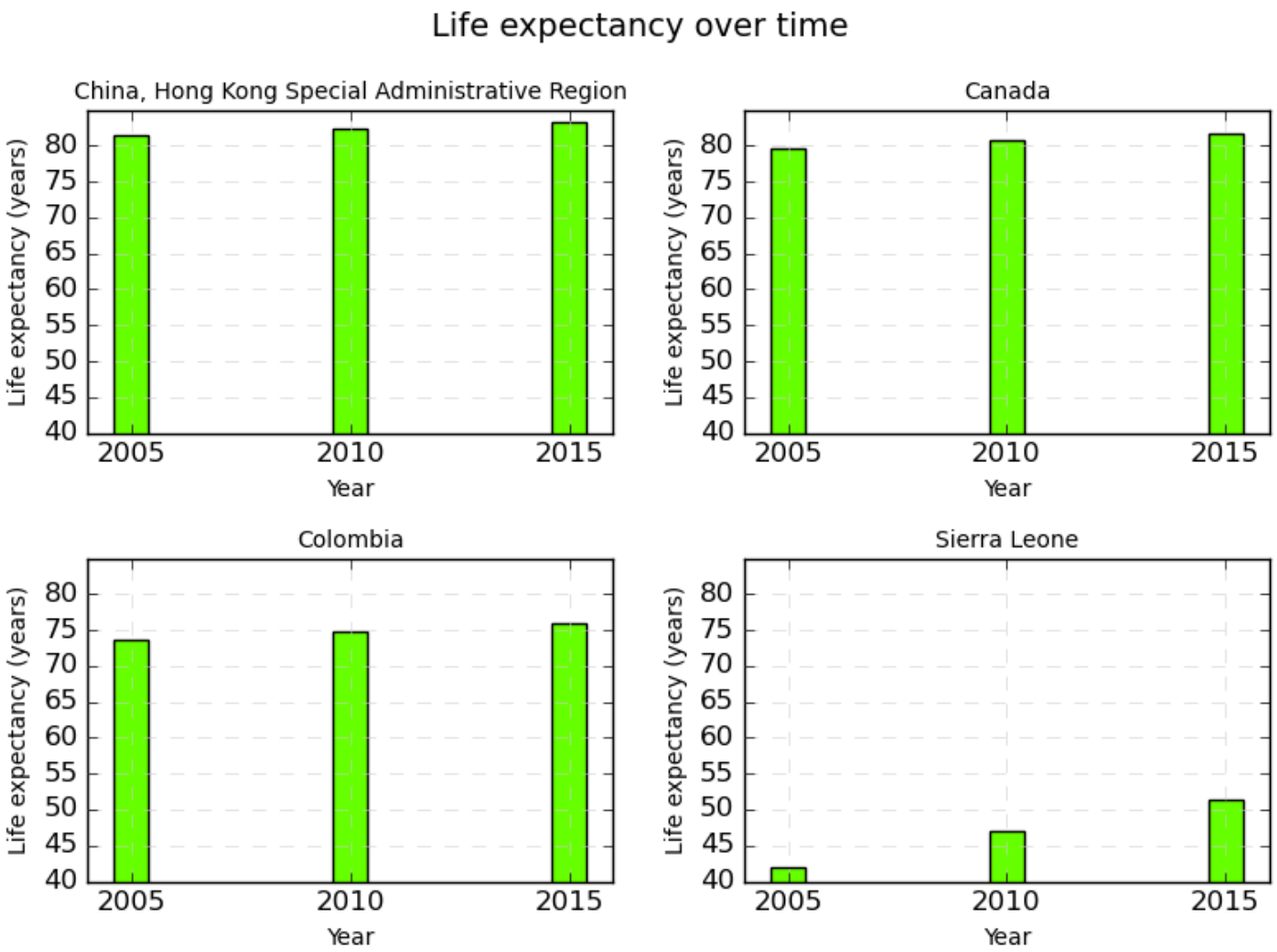
```
---
Countries with life expectancy over 80 years:

    Country or Area                                    Year
    Australia                                          2005    80.336
                                                       2010    81.476
                                                       2015    82.363
    Austria                                            2010    80.130
                                                       2015    80.986
    Belgium                                            2015    80.524
    Canada                                             2010    80.761
                                                       2015    81.787
    China, Hong Kong Special Administrative Region     2005    81.407
                                                       2010    82.428
                                                       2015    83.404
    China, Macao Special Administrative Region         2005    80.970
                                                       2010    82.069
                                                       2015    83.285
    Denmark                                            2015    80.143
    Finland                                            2015    80.701
    France                                             2010    80.952
                                                       2015    81.939
    Germany                                            2015    80.445
    Greece                                             2010    80.008
```

Masking operation

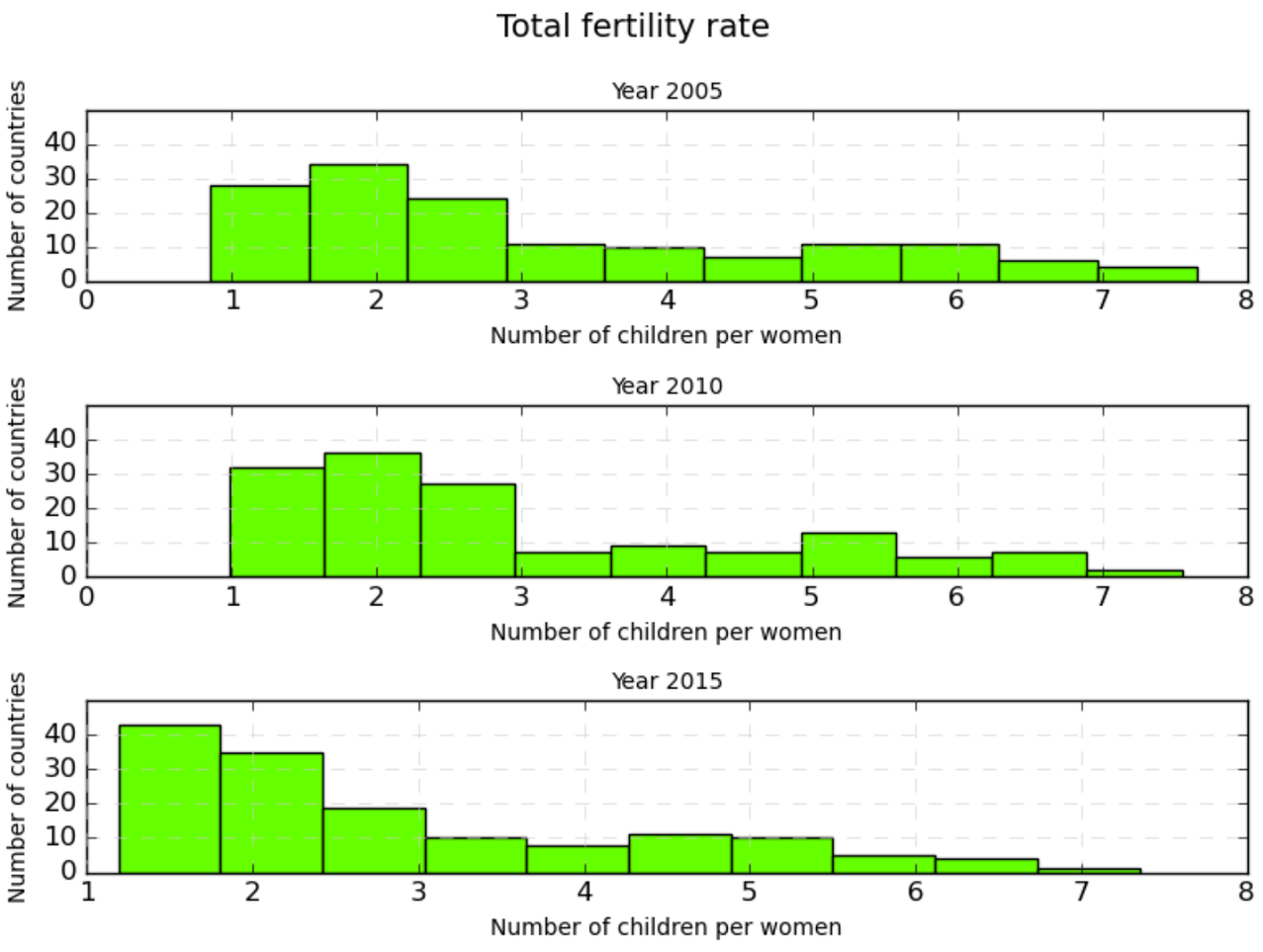
Plus, the entire data set, including two manually added and pivoted columns:



And finally shows the trends for the life expectancy of a few selected countries:



And the evolution in fertility rate in all countries



Above and beyond the minimum requirements

While working on the assignment we encountered a problem where code ran on Kevin's machine but not on Tyson's machine.

After digging into the causes, we discovered that Tyson was using a different version of Pandas than Kevin. We installed the same version of Pandas on both machines and then everything worked well.

However, in real life this scenario is important to avoid, and for this we used two wide-spread techniques:

- Packaging the project with Nix: https://nixos.org/. This way anyone on an x86_64-linux system can run the project with: **nix run**

```
[kamadorueda@nixos:/data/secrets/u/2022-spring/ensf-592/final-project-kernel-panic]$ nix run
---
Note: this program requires the following dependencies:
```

- Building a container so that someone else can run it with Docker (https://www.docker.com/) or Podman (https://podman.io/) or any other container runtime:

```
[kamadorueda@nixos:/data/secrets/u/2022-spring/ensf-592/final-project-kernel-panic]$ nix build .#container
[kamadorueda@nixos:/data/secrets/u/2022-spring/ensf-592/final-project-kernel-panic]$ docker load < result
Loaded image: kernel-panic:latest
[kamadorueda@nixos:/data/secrets/u/2022-spring/ensf-592/final-project-kernel-panic]$ docker run --interactive --tty --volume "$PWD:/data" kernel-panic
---
Note: this program requires the following dependencies:
    matplotlib==3.5.1
```