Correlation of GDP and fertility rate

Data Collection

What data will you collect or create?

The project uses data from two external datasets:

1) GDP per Capita, retrieved from https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

Format: Comma Separated Values (CSV)

Location in repository: data/raw/API_NY.GDP.PCAP.PP.CD_DS2_en_csv_v2_2183937.csv

Size: 184 kb

Data description: This dataset has GDP per capita data for almost all countries and regions over the last 60 years for each year.

This dataset comes from a very reliable source, The World Bank, and that was the main reason why we used it as the dataset for our project. It contains GDP per capita data of each country for many years and it was easy to choose the year with the most relevant data to present the correlation between the fertility rate. Since we work in Python3, which has some well known libraries for reading and parsing CSV files, we decided to save the dataset in CSV format.

2) Fertility rate(births per woman), retrieved from https://data.worldbank.org/indicator/SP.DYN.TFRT.IN

Format: Comma Separated Values (CSV)

Location in repository: data/raw/API_SP.DYN.TFRT.IN_DS2_EN_csv_v2_2163843.csv

Size: 168 kb

Data description: This dataset has Fertility rate data for almost all countries and regions over the last 60 years for each year

Same as the previous dataset, this dataset also comes from The World Bank, it is of the same format and data is almost identically organised within the file as the GDP data in the previously mentioned dataset. We could easily reuse both datasets and present the correlation for any year of our choice from 1960 to 2019. Same as the above mentioned reasons, we also decided to save this dataset in the CSV format.

We created three artifacts:

1) Python code saved in two jupyter notebooks files located in the notebooks directory:

Digital Object Identifier: 10.5281/zenodo.4698443

- 01_data-preprocessing.ipynb
- 02_data-vizualization.ipynb

The first notebook parses the raw data from two external datasets and merges them together for the year 2016. During the process, some countries that don't have data for 2016 are omitted from the merged dataset. The second notebook uses the merged data to create a scatter plot that shows the correlation between the GDP per capita and fertility rate for all countries and regions of the merged dataset.

2) Merged dataset created after processing raw data from two external sources

Digital Object Identifier: 10.5281/zenodo.4698406

Format: Comma Separated Values (CSV)

Location in repository: data/processed/gdp_fertility.csv

Size: 9 kb

Description: This file contains GDP per capita and fertility rate of all countries that have that information for the year 2016. The data from this file will be used later for the graphical representation of our experiment

3) Graphical representation of the merged dataset

Digital Object Identifier: 10.5281/zenodo.4698428

Format: Portable Graphics Format (png)

Location in repository: reports/figures.correlation.png

Size: 21kb

Description: This is a scatter plot of the merged data that clearly shows the correlation between the fertility rate and the GDP per capita for the year 2016.

How will the data be collected or created?

The two external datasets are both going to be downloaded from the data.worldbank.org(exact URLs are already specified above), in the csv format. Those two datasets are going to be parsed in a same way, because their structure is identical. Since we want to present our research for the year 2016, we will extract only the values of the 2016 column, along with the country names. The countries without value aren't considered. To create our merged dataset, We are going to group two parsed datasets by country name and save it in the csv format. Using matplotlib, we will present the merged dataset as a scatter plot with GDP as x axis and fertility rate as y axis.

The structure of our project is following:

There are four directories: data, notebooks, reports and documentation . The data directiory consists of two subdirectories, one for raw data which is collected from the external sources, and one for processed data, which is generated after processing the raw data. In the notebooks directory, the jupyter notebooks are stored. They should be named as the combination of number and its purpose, and they should be executed in order specified by the number. The reports directory contains generated visualizations of the processed data. The documentation directory contains metadata files, and data flow chart. The project is stored and versioned under the GitHub repository: https://github.com/jpetrovi/GDP_Fertility_Correlation.

Documentation and Metadata

What documentation and metadata will accompany the data?

There will be a README file in the GitHub repository that clearly and in detail describes the required processes and steps in order to reproduce the experiment. Additionally, a data flow chart will be included to visualize the whole process of the experiment and a brief description of the experiment. Those documents should contribute to better understanding of the experiment procedures and segments.

There are three xml metadata files that describe the project and generated data, all written in well known Dublin Core Metadata Initiative (DCMI) standard:

- metadata.xml - File that contains general information about the project
- correlation_csv_metadata.xml - File that contains information about the generated csv file
- correlation_plot_metadata.xml - File that contains information about the scatter plot of generated data of the csv file

Ethics and Legal Compliance

How will you manage any ethical issues?

No particular ethical issue is foreseen with the data to be used or produced by the project and the data used isn't supposed to do any harm to any group or individual. Both external dataset have the Creative Commons Attribution 4.0 International license which allows users to copy, modify and distribute data in any format for any purpose, including commercial use.

How will you manage copyright and Intellectual Property Rights (IPR) issues?

The owner of the created data is Jovan Petrovic. The generated data will be licensed under Creative Commons Attribution 4.0 International license and will be available for any kind of reuse by third parties. The source code is stored on the public GitHub repository https://github.com/jpetrovi/GDP_Fertility_Correlation and is available for any kind of reuse by third parties

Storage and Backup

How will the data be stored and backed up during the research?

The project is tracked by a version control tool GitHub that tracks changes and maintains project versions. If there is a need to check the previous state of some project data, it could be easily accessible with GitHub. Since the project data doesn't consume much memory, there is no need for additional storage services. Our research team is going to be responsible for backup of data and the recovery in case of some major incident. The backup will be performed weekly by our team and data and code will be stored on our private Google Cloud instance. That way, we will ensure that our data is safely stored on two external hosting servers, Google Cloud and GitHub and in case of some major incident on one of those two servers, we will just load the data from the working server and it will never be lost.

How will you manage access and security?

Since our project is stored on a public GitHub repository, there are no access restrictions to the source code and created data, anyone who wants may reuse our project or if someone wants to collaborate with us, we are open for pull request that will be reviewed by our team. GitHub has already mechanisms that ensure data security and prevents unauthorized users from altering the project content.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

None of our generated artifacts doesn't have to be destroyed and as already mentioned, it is free and available to anyone interested in it. Other than the data published in the repository, no other data will be generated or saved. This research should serve as a clear indicator that countries and regions with poor economic standard have more children per woman than countries with better economic situation and this research could be further used in many other research disciplines such as sociology, anthropology and economy. Currently there is no time limit that should indicate how long the data should be stored and be publicly available and as long as that is the case, it will be present for any kind of reuse.

What is the long-term preservation plan for the dataset?

As already mentioned, the data will be stored in a public GitHub repository, and there will be no costs or charges for the storage.

Data Sharing

How will you share the data?

This DMP is going to be published on Zenodo, which is probably the biggest open access repository for any kind of research and from there, the interested users will be able to find out how and where to find the data and source code located in public GitHub repository. The data is already available in the repository and produced data and source code will obtain a Digital Object Identifier (DOI) that serves as a permanent identifier on the network.

Are any restrictions on data sharing required?

There are no data sharing restrictions present in our project, since it is publicly available to any interested parties.

Responsibilities and Resources

Who will be responsible for data management?

The owner and writer of this DMP, as well as responsible for all data management activities is Jovan Petrovic and it will be reviewed by the teaching staff of Vienna University of Technology.

What resources will you require to deliver your plan?

We used Jupyter Notebook for implementation of our source code in python, together with some well known python libraries such as matplotlib, pandas, etc. The code will be stored and versioned in GitHub. There is no additional hardware and software required for storage of data and implementation of project because this project doesn't consume much memory and it doesn't require any special equipment for its conducting.