Correlating global temperature change on the number of children born in Vienna

A Data Management Plan created using DMPonline

Creator: Aleksandar Pavlovic

Affiliation: Other

ORCID iD: 0000-0001-6887-9515

Template: European Commission (Horizon 2020)

Project abstract:

The aim of this pseudo experiment is to analyze the relation between the mean near surface temperature deviation and the number of children born alive in Vienna. The experiment is conducted, by means of visual data science (exploring the data by plotting it in an appropriate manner).

Last modified: 22-04-2019

Correlating global temperature change on the number of children born in Vienna - Detailed DMP

1. Data summary

State the purpose of the data collection/generation

The purpose of the data collection is to gather open data information about the mean near temperature deviation and the number of live-births in Vienna, such that the research question, if there is a relation between those two entities, can be answered. The datasets should ideally cover the years from 2002 to 2017, such that some reasonable data analysis about the current evolvement can be performed.

Explain the relation to the objectives of the project

The collected data should help answer the research question, if there is a relation between mean near surface temperature deviation and number of live-births in Vienna, by aggregating it over a certain time unit (e.g. year) and plotting the time series in line plots and scatter plots. Furthermore the correlation of the two variables will be measured in one value by calculating the Pearson correlation coefficient.

Specify the types and formats of data generated/collected

The following two datasets are used for the experiment:

Number of live-births in Vienna

The dataset was downloaded from the Open Data Austria website (https://www.data.gv.at/), accessed via https://www.data.gv.at/katalog/dataset/f54e6828-3d75-4a82-89cb-23c58057bad4 on April 19, 2019 It is freely available under the Creative Commons (CC BY 4.0) licence and has the unique id f54e6828-3d75-4a82-89cb-23c58057bad4.

It contains data about the number of live briths in Vienna from 2002 to 2017 group by age of the mother and gender of the child and is a comma separated file.

Mean near surface temperature deviation

The dataset was downloaded from the EU Open Data website, accessed viahttps://data.europa.eu/euodp/en/data/dataset/zQAEvhkR7H0NQYU1HP5fA on April 19, 2019 It is freely available under the Creative Commons (CC BY 4.0) licence and has the unique id cli_iad_td.

It contains data about the global mean near surface temperature deviation from 1850 to 2017 from different sources and is a tabular separated file. The source "TD_GLB,NOAAGLOBALTEMP,DEGC" will be used for the purpose of the experiments.

Both datasets are stored in standard widely used open formats and should thus be easily reusable.

The following data is generated during the experiment:

Processed dataset (DOI: 10.5281/zenodo.2648703)

This dataset contains three variables, the year of the measurement, the number live-births in Vienna during the respective year and the "TD_GLB,NOAAGLOBALTEMP,DEGC" near surface temperature deviation. The data ranges from 2002 to 2017.

It is a comma separated file to again allow for easy reusability and to stay consistent with the formats of the previously described datasets.

Plots (DOI: 10.5281/zenodo.2648779)

- Time series line plot
  - The x-axis depicts the years of the sampled data and ranges from 2002 to 2017
  - There are two y-axises:
    - the left one depicts the global temperature deviations and belongs to the blue curve (Temp dev)
    - the right one depicts the number of live-births in Vienna and belongs to the red curve (Live-births)
- Scatter plot (project/plots/scatter-plot-with-linear-model.jpg):
  - The x-axis depicts the global near surface temperature deviation
  - The y-axis depicts the number of live-births in Vienna
  - The black dots represent data points
  - The blue line shows a linear regression model fitted to the data
  - the grey area guarding the blue prediction line depicts the 95% confidence interval of the respective prediction

Both images are saved in the JPEG format, as it allows for easy interoperability on different OS systems and does not consume too much memory.

Report (DOI: 10.5281/zenodo.2648785)

Knitting the report.Rmd file creates a report.pdf file, that shows the main results together with the corresponding code for easy understandability of the experiment. Since this report is intended only for human agents, who want to reproduce the experiment by following every step, pdf was chosen as the format of choice. Furthermore pdf is widely adopted by the research and industrial community for human centered docuemnts and is thus the format of choice.

Specify if existing data is being re-used (if any)

As described in the previous section both datasets (*Number of live-births in Vienna* and M*ean near surface temperature deviation*) are provided by the Open Data Austria and EU Open Data website under CC BY 4.0 licence, thus allowing the reuse of the data.

Specify the origin of the data

The m*ean near surface temperature deviation dataset* is provided by the EEA and was published by Eurostat.
The *number of live-births in Vienna dataset was published by the city of Vienna.*

State the expected size of the data (if known)

The size of the temperature dataset is about 2KB and the size of the number of live-births in Vienna dataset about 38KB.
The size of the processed dataset is about 1KB, since most information that is not necessary for the experiment is discarded. No additional costs are needed for storing and sharing the data, due to the very small file sizes.

Outline the data utility: to whom will it be useful

The data is shared by depositing it in a public GitHub repository (https://github.com/AleksVap/Effect-GlobalTemperature-NumberOfBirthsVienna, DOI: 10.5281/zenodo.2648771) together with all the source code needed to generate the experiment results.
It will be available to all users, who can access a public GitHub repository.
It might be useful to anyone who wants to discover patterns between the total number of births in Vienna and the mean near surface temperature deviation and is stored in a csv file to allow for easy reusability.
One may use the source code to process new versions of the external datasets.

2.1 Making data findable, including provisions for metadata [FAIR data]

Outline the discoverability of data (metadata provision)

In the previously mentioned GitHub repository a folder called documentation will be created, containing information about the project.
There will be a metadata.xml file, stating the author of the project, the project type, date, coverage, rights and so on.
Furthermore there will be a documentation.txt file outlining the results of the experiment (information about the created dataset e.g. its header and information about the plots e.g. how to read the axes).
Moreover there will be an architecture.png file outlining the architechture of the experiment, to outline briefly what kind of data is produced in what way.

Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?

To allow for data citation as well as identification a DOI will be assigned to every result of the experiment and the scripts used to obtain them in the github repository.

Outline naming conventions used

Filenames consisting of different words are separated by capitalizing the first letter of every new word, for instance a "time series line plot" is named "timeSeriesLinePlot.jpg".

Outline the approach towards search keyword

Since GitHub already provides functionality to search for repositories using key words, no further measure is taken.

Outline the approach for clear versioning

GitHub is used as the version control systems (VCS) of choice. Every change of the content of the repository is explained using commit messages.

Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how

A metadata.xml file will be stored at the documentation folder containing the following information:

- project title
- creator
- subject
- project description
- date of the experiment
- project type
- script formats
- data sources (and ids)
- language of the project
- data coverage (years)
- rights for resuse

2.2 Making data openly accessible [FAIR data]

Specify which data will be made openly available? If some data is kept closed provide rationale for doing so

All data and code available in the project repository (including scripts, reports, graphs and data) will be made available to the public (free of charge).

Specify how the data will be made available

The data will be made available by placing it into a public Github repository https://github.com/AleksVap/Effect-GlobalTemperature-NumberOfBirthsVienna (with DOI: 10.5281/zenodo.2648771). It will be made accessible during the whole project life cycle and will be available to the public (everyone who can access a Github repository) free of charge.
Github should place minimal restrictions on the availability of the project data, as it is widely adopted and does not place any additional repository costs as long as the project is smaller than 1GB, which is in this case easily satisfied.

Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

To perform the experiment certain software is needed including *RStudio (version 1.1.456), MiKTeX* (*version 2.9) and R (version 3.5.1).*
Furthermore the following libraries are needed:

- *rmarkdown - version 1.10*
- *ggplot2 - version 3.0.0*

The README.txt file lists all these libraries and software packages needed to execute the given scripts. Since these software tools are all very well known and since installing libraries in R is only one line of code, no further information about how to install the respective software is provided.

Specify where the data and associated metadata, documentation and code are deposited

Project structure:

- documentation - Contains the documentation of the results (documentation.txt), the project metadata (metadata.xml) and the experiment architecture (architecture.png).
- src - Contains the code for preprocessing the data and conducting the experiment plus the results (see following subfolders)
- src/data/raw - Contains the downloaded (raw) datasets
- src/data/processed - Contains the preprocessed and merged dataset
- src/preprocessData.R - Contains the code for preprocessing the data.
- src/report.Rmd - Contains the code necessary to load the preprocessed data and to execute the experiment and is used to generate the report.pdf file
- src/plots - Contains the result plots of the experiment

Specify how access will be provided in case there are any restrictions

Access will be provided through placing the project into a Github repository (https://github.com/AleksVap/Effect-GlobalTemperature-NumberOfBirthsVienna, DOI: 10.5281/zenodo.2648771), no restrictions (other than access to Github, which is free of charge) are imposed on the accessibility of the project.

2.3 Making data interoperable [FAIR data]

Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.

Interoperability is facilitated by using open standards like CSV, TSV, JPG and PDF for the results of the experiment (data, graphs and reports). Furthermore open and widely adopted software (R and LaTex) will be used to make the experiments interoperable.
Moreover the metadata.xml file uses the Dublin Core vocabulary, which is commonly used for metadata specification.

Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

No specific vocabulary will be used to indicate the datatypes of the columns, since the resulting dataset is only very small (16 rows and 3 columns) and since the correct datatypes are inferred by many software packages during loading automatically (e.g. by R).

2.4 Increase data re-use (through clarifying licenses) [FAIR data]

Specify how the data will be licenced to permit the widest reuse possible

The MIT lincence will be used for the data, as it permits the use, copy, modification, publishing, distribution, sublicensing and selling of the data free of charge.

Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed

The data is made available starting from April 21, 2019 to the public without any embargo.

Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why

Since the whole project is licenced using the MIT licence, no restriction on the usage of any part of the project by third parties is imposed.

Describe data quality assurance processes

The data quality is ensured through multiple means, first by using reliable well established data publishers (the city of Vienna and Eurostat), next the data was checked visually for any significant outliers and in the case of the temperature deviation, multiple sources were compared (all stated in the same dataset), which differed marginally in absulote value, but shared the same global trend.

To extend the data quality (and reusability) of the result (processed data and plots) further, a document.txt file contains descriptions and explanation of what can be seen in the visualisations and in the processed dataset.

Specify the length of time for which the data will remain re-usable

The project should remain usable for the next two to five years, afterwards the data would already be outdated and thus presumably irrelevant, if not extended by new data.

3. Allocation of resources

Estimate the costs for making your data FAIR. Describe how you intend to cover these costs

No costs are necessary to make the project FAIR, as the used public repository is free of charge (for the small size of the project) and as no commercial software was needed to create any of the results (merely open software and standards).

Clearly identify responsibilities for data management in your project

All aspects of the project were carried out by Aleksandar Pavlovic (orcid: 0000-0001-6887-9515), who is thus responsible for the entire data management plan.

Describe costs and potential value of long term preservation

No costs are planed for the long time preservation of the data, as the whole project is only very small (way below 1GB) and GitHub allows free sharing of projects that do not exceed 1GB of memory.
Long term preservation of this project has only a value for a few years (2 to 5), as afterwards the experiment results will be very outdated and thus presumably be of little interest to the public. No update of the raw data is planed for the future.

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

The data will be stored and versioned using GitHub, allowing for a secure storage and transfer of the data through the following means.
On the one hand GitHub audtis a lot of actions, like logins, password resets repository access, two-factor authetication requests and et cetera. Furthermore it provides means of authentication, signing of code commits and acces over SSH and HTTPS, to ensure that the data stays secure.
Moreover older version of the projects can be accessed easily by checking out previous commits of the project.

5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

The processed data does not contain any information about one indivual person. The data is aggregated over a whole year and thus does not need to take any measures to protect the identification of certain individuals.
Also the used external datasets are publically available and were annonymized by the institution, that published them accordingly (e.g. by removing the district of birth of the babies etc.).
Furthermore this yields, that no ethical measure need to be taken during the save or transfer of the data.
Since the external datasets were published using a CC BY 4 licence, no ethical issues can be seen on the reuse of the data.

6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

No additional procedures by any means were established for the data management plan.