Comparison of divorce rate between Austria and the EU

1. Data Summary

State the purpose of the data collection/generation

No new data was collected for this project. Only data from the Austrian and European open data portals has been used. The generated data are images that are used to draw conclusions between the divorce rate of Austria and the EU28.

Explain the relation to the objectives of the project

The input data contains the annual number of divorces for Austria and the EU.
The produced data is a visual analysis of the correlation between the Austrian and the EU28 divorce rate.

Specify the types and formats of data generated/collected

Title: Ehescheidungen (Statistik Austria)
Description: Divorce statistic of austria
Type: Dataset
Format: csv
Source: Input

Title: Divorces by duration of marriage (Eurostat)
Description: Divorce statistic of the EU
Type: Dataset
Format: tsv
Source: Input

Title: scatter
Description: Scatterplot
Type: Image
Format: png
Source: Produced

Title: time_change
Description: Time Change Plot
Type: Image
Format: png
Source: Produced

Title: time_corr
Description: Time Correlation Plot
Type: Image
Format: png
Source: Produced

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

Ehescheidungen (Statistik Austria) and Divorces by duration of marriage (Eurostat) are re-used for this project.

Specify the origin of the data

Ehescheidungen (Statistik Austria) was downloaded from https://www.data.gv.at/katalog/dataset/2d8ad82c-4730-3354-9971-9406f2ccf72c.
Divorces by duration of marriage (Eurostat) was downloaded from https://ec.europa.eu/eurostat/web/products-datasets/-/demo_ndivdur.
scatter was produced during research.
time_change was produced during research.
time_corr was produced during research.

State the expected size of the data (if known)

Ehescheidungen (Statistik Austria) has a size of 1KB
Divorces by duration of marriage (Eurostat) has a size of 297KB
scatter has a size of 15KB

time_change has a size of 30KB
time_corr has a size of 15KB

Outline the data utility: to whom will it be useful

Storing the data makes it easy for other people to rerun and validate the correctness of the experiment. It also avoids the possible loss of the original data through the third-party providers and guarantees access to the data.

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

Outline the discoverability of data (metadata provision)

The metadata description for Ehescheidungen (Statistik Austria) can be found in the jupyter notebook file "divorce_analysis.ipynb" or can be accessed online at https://www.data.gv.at/katalog/dataset/2d8ad82c-4730-3354-9971-9406f2ccf72c. The metadata description for Divorces by duration of marriage (Eurostat) is stored in the same file "divorce_analysis.ipynb" or online at http://data.europa.eu/euodp/en/data/dataset/bRJAS74ZDdIpeU7mnKhMiA.

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?

The original providers of the input dataset only use internal identifiers for their dataset. The Github repository however has assigned a DOI (10.5281/zenodo.2647476).

Outline naming conventions used

The input files are saved with their original file names as provided by their original source.
Other files, code and data use naming conventions commonly used with python (https://visualgit.readthedocs.io/en/latest/pages/naming_convention.html)

Outline the approach towards search keyword

Not implemented.

Outline the approach for clear versioning

The original release was versioned as v1.0.0. Every further release however will follow a date based versioning pattern <year>.<month>.<day>. <sequence_within_day> (e.g. 2019.04.19.01).
Since the project is hosted as a git repository, keeping track of changes to the data is handled automatically.

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

The metadata in the jupyter notebook contains information about the original fields of the input data and what part of it actually is used in the project.

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

Ehescheidungen (Statistik Austria) is open access.
Divorces by duration of marriage (Eurostat) is open access.
scatter is open access
time_change is open access

time_corr is open access

Specify how the data will be made available

The data can be accessed through Github (https://github.com/martinpichler/data_stewardship_ex1).

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)?

The data can be downloaded either from Github directly (single files or zip) or can be downloaded with git.
The input data files are stored as .csv and .tsv files and need no additional software.
The produced data are .png images and an image viewer is needed to open them.
The source code of the experiment is contained within a jupyter notebook. A working python installation is needed to run it. The git repository contains a readme.md that describes how to run it and which dependencies the notebook has. Alternatively, docker can be used to run the experiment (also described in the readme.md).

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

As stated in previous sections, the complete project is available on Github (https://github.com/martinpichler/data_stewardship_ex1)

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

No restrictions, access is public and open to everyone.

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.

All the data used and produced is based on non-restricted or proprietary data formats (e.g. .csv, .tsv, png, python, jupyter, git).
There is no domain-specific vocabulary to take in consideration.

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?

Standard vocabulary is used, no mapping required.

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 project is licenced under the MIT licence (https://opensource.org/licenses/MIT) and open for re-use by anyone.

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

Already openly available.

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

No restrictions, data and source code can be used by anyone.

Describe data quality assurance processes

The project only uses already available data and no new data will be generated in the future. The used data was checked for completeness and correctness.

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

There is no limit on how long the data will remain re-usable.

3. Allocation of resources

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

Since the data is stored in a public Github repository, no additional cost has to be covered.

Clearly identify responsibilities for data management in your project

Martin Pichler
mpichler.dev@gmail.com
https://orcid.org/0000-0001-5305-9063

Describe costs and potential value of long term preservation

Since the data is stored in a public Github repository, no additional cost has to be covered.

4. Data security

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

Data is stored in a git repository on Github and can only be modified by the owner or an administrator. If data is lost locally, the original data can be downloaded from the Github repository. No sensitive data is stored and thus the repository does not need to be addressed.

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

No ethical issues exist

6. Other

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

No additional procedures are used.