Plan Overview

A Data Management Plan created using DMPonline

Title: Analysis of the distribution of inhabitants of Austria by altitude

Creator: David Wagner

Affiliation: Other

ORCID iD: 0000-0002-5142-1054

Template: DCC Template

Project abstract:

In order to get a better understanding of the altitude on which the Austrians live certain data had to be collected and transformed. At this point in time the analysis was only done at the "Bezik" level of Austria. The final result is that ever Bezirk in Austria was connected with the number of inhabitants in this Bezirk and furthermore the altitue of the main city in the bezirk was added.

ID: 39613

Last modified: 22-04-2019

Analysis of the distribution of inhabitants of Austria by altitude

Data Collection

What data will you collect or create?

Data collected and reused:

- Population Data of Austria per Bezirk
  - https://www.data.gv.at/katalog/dataset/3bfba412-7053-3a60-937a-8c3dd2c71294
    - Unique ID of the data on data.gv.at: 3bfba412-7053-3a60-937a-8c3dd2c71294
  - Three CSV files:
    1. Bezirk Name mapping to internal code
       1. Name: OGD_f0743_VZ_HIS_GEM_4_C-GRGEMAKT-0
       2. Size: 7 Kb
       3. Location: ./data/raw
    2. Year mapping to internal Code
       1. Name: OGD_f0743_VZ_HIS_GEM_4
       2. Size: 8 Kb
       3. Location: ./data/raw
    3. Actual Population data
       1. Name: OGD_f0743_VZ_HIS_GEM_4
       2. Size: 40Kb
       3. Location: ./data/raw
- Altitude Data
  - from DBpedia
    - Accessed directly and stored in a pandas DataFrame
  - Missing values that were not found on DBpedia
    - Name: MissingHM
    - Location: ./data/raw
    - Generated manually
    - Format: CSV
      - Non proprietary
    - Size: <2 Kb

Data created:

- Comined table of Population of Austria in 2011 per Bezirk plus the altitude of the main city in the Bezirk.
  - Name: Population2011_altitude
  - Location: ./data
  - Format: CSV
    - Non proprietary
  - Size: <9 Kb
  - DOI: https://doi.org/10.5281/zenodo.2648511

How will the data be collected or created?

Collected:

- Population Data of Austria per Bezirk
    - Can be found on data.gv.at
        - Manual download
- Altitude Data:
    - Can be queried from DBpedia
    - Data that can't be found on DBpedia
        - Manual lookup on Wikipedia

Created:

- Merge of the data that was found and queried
- Stored in a Pandas DataFrame and later on exported as CSV

Standards that are used:

- CSV

Folderstructure:

- Raw data is stored in: ./data/raw
- Generated data is stored in: ./data

Versioning:

- Raw data can be replaced if new data for more recent years will be available
    - The format has to be the same and should include the old data - as it is the default right now
- Generated data: Can be generated for any desired year and saved as a CSV with the included code; There is no real versioning required

Naming:

- The default Names that can were used by data.gv.at were kept.

Documentation and Metadata

What documentation and metadata will accompany the data?

The experiments were run using Jupyter Notebooks. Two of those exist:

1. The first one is used in order to preprocess the raw data and transform and merge it

    1. Location: ./01-PopulationAltitude_Preprocessing.ipynb
2. The second one is used to generate plots that aim to answer the initial research question
    1. Location: ./02-PopulationAltitude_Plots.ipynb

On top of this a readme file in Markdown format will be included that will walk useres through all the steps of running the Jupyter Notebooks.

- This file will be stored in: ./README.md

Lastly a matadata file in xml format will also be included

- This file will be stored in: ./documentation/metadata.xml

Ethics and Legal Compliance

How will you manage any ethical issues?

No personal data is used for the experiment, so there are no real ethical issues that could arise.

- This also means that there is not sensible data that needs to be anonymised.

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

Data from data.gv.at:

- This is the main data source.
- Everything is under: Creative Commons Attribution License 3.0

Data from DBpedia:

- This is the second main data source where the altitude of cities is queried from.
- Everything is licensed under: Creative Commons Attribution-ShareAlike 3.0 License and the GNU Free Documentation License

Therefore there are not issues with reusing the raw data as well as the data that is produced by the experiment.

Storage and Backup

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

Storage:

As the total size of all files is less than 1 Mb storage is not an issue.

Backup:

The data is stored locally for ease of use, but it is also saved on University Servers of TU Vienna.

- Backups are only made when the data changes, which does not happen often as we are dealing with altitude data of cities - both of with do not really change a lot over time. The only attribute that is prone to change is the number of inhabitants, but also this only happens every few years.

Recovery:

In case of fatal issues where the data is lost locally it will still be possible to access it from the original sources (data.gv.at and DBpedia).

How will you manage access and security?

Risks to data security:

- As we are not dealing with any personal data there is little risk of personal data being published.
- However, there is still the risk that data could be changed or deleted. However this is also of little concern since all the data that was used is publically available so it can easily be checked it anythin was changed.
    - Chrosschecking with the original data sources is also a good option to continously check the health of the data.

Access to the data:

- Access to the data (and the backup) will only be granted for the people that also conducted the experiment.

The generated data was published on Zenodo:

- https://doi.org/10.5281/zenodo.2648511

Safe transfer of data:

- Will be granted by using the VPN network of the Technical University of Vienna.

Selection and Preservation

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

Which data should be preserved?

- Both Jupyter Notebooks
    - ./01-PopulationAltitude_Preprocessing.ipynb
    - ./02-PopulationAltitude_Plots.ipynb
- ./README.md
- ./documentation/metadata.xml
- ./requirements.txt
- ./documentation/description.txt
- ./documentation/architecture.png
- ./data/MissingHM.csv
- The ouput data also does not necessarily have to be preserved, as it can easily be created with a Jupyter Notebook and the iput data. Howeverm in order to make future research easier and enable other researchers to find the data that was already enriched and extended it was stored
    - https://doi.org/10.5281/zenodo.2648511

Which data is not required to be preserved?

- The input data is not required to be preserved, as it can easily be restored by accessing the data.gv.at homepage

There are not possible contractual or lega issues with storing the data.

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

The long term preservation will be handled using digital perservation services like Zenodo. The code will furthermore be stored on Github.

As these services are free, there are no expected costs.

All the data is ready to be shared by default, as there is no anonymazation or similar efforts required.

Data Sharing

How will you share the data?

The output data will be stored in the following repository:

- Zenodo
    - DOI: https://doi.org/10.5281/zenodo.2648511
    - License: Creative Commons Attribution 4.0 International
    - Availabe since: 22nd of April 2019

The Jupyter Notebook, as well as all the documentation, input files and output were pubished on:

- Github with the integration to Zenodo
    - DOI: https://doi.org/10.5281/zenodo.2648637
    - License: MIT License
    - Availabe since: 22nd of April 2019

Are any restrictions on data sharing required?

Due to the licesnses mentioned above there are not real restrictions on sharing and reusing the data.
There is also no exlcusive use of the data.

Responsibilities and Resources

Who will be responsible for data management?

Main responsibility lies with David Wagner, as he is the only person conducting this experiment.

What resources will you require to deliver your plan?

There is no training needed, since the key people in the experiment had a great Data Stewardship lecture at their university.
The only additional software that is required are the data and code repositories:

- Github
- Zenodo

Both do not charge anything for this project.