Project No. 957117

Project acronym: EVIDENT

Project Title:

bEhaVioral Insgihts anD Effective eNergy policy acTions

Deliverable 5.4

Updated data documentation

Programme: H2020-LC-SC3-EE-2020-1
Start date of project: December 01, 2020
Duration: 36 months

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957117

Document Control Page

Deliverable Name	Updated data documentation
Deliverable Number	D5.2
Work Package	WP5
Associated Task	Task 5.2
Covered Period	M18-M30
Due Date	May 31, 2023
Completion Date	May 31, 2023
Submission Date	May 31, 2023
Deliverable Lead Partner	CERTH
Deliverable Author(s)	Georgios Sidiras (CERTH), Tilemahos Efthimiadis (JRC), Grigorios Kountatos (PPC), Panagiotis Sarigiannidis (UOWM), Anna Triantafyllou (UOWM), Athanasios Liatifis (UOWM), Georgios Fragoulis (UOWM), Peter Rosengren (CW), Paris Karypidis (DUTH)
Version	V1.0

Dissemination Level		
PU	Public	**X**
CO	Confidential to a group specified by the consortium (including the Commission Services)	

Document History

Version	Date	Change History	Author(s)	Organisation
0.1	April 3, 2023	Table of contents, Initial version	Georgios Sidiras	CERTH
0.2	April 28, 2023	Section 3	Paris Karypidis, Grigorios Kountatos	DUTH, PPC
0.3	May 5, 2023	Section 2	Panagiotis Sarigiannidis, Anna Triantafyllou, Athanasios Liatifis, Georgios Fragoulis	UOWM
0.4	May 10, 2023	Inclusion of chapter on ethics. General improvements.	Tilemahos Efthimiadis	JRC
0.5	May 11, 2023	Section 3 and section 2 updates	George Sidiras	CERTH

0.6	May 16, 2023	Updates on section 3	George Sidiras	CERTH
0.7	May 19, 2023	Version for internal review	George Sidiras, Tilemahos Efthimiadis	CERTH, JRC
0.8	May 26, 2023	Internal Review	Paris Karypidis, Tilemahos Efthimiadis	DUTH, JRC
0.9	May 29, 2023	Quality and Risk Review	Dimosthenis Ioannidis	CERTH
1.0	May 31, 2023	Final version	George Sidiras, Paris Karypidis, Tilemahos Efthimiadis	CERTH, DUTH, JRC

Internal Review History

Name	Institution	Date
Paris Karypidis	DUTH	May 26, 2023
Tilemahos Efthimiadis	JRC	May 26, 2023

Quality Manager Revision

Name	**Institution**	**Date**
Dimosthenis Ioannidis	CERTH	May 29, 2023

Table of Contents

List of Figures

List of Tables

Acronyms

Acronym	Explanation
API	Application Programming Interface
AI	Artificial Intelligence
APB	Average price bias
CW	Check Watt
CC	Creative Commons
DX.Y	Deliverable X.Y
DOI	Digital object identifier
DCE	Discrete choice experiment
DSO	Distribution system operator
EPC	Ethics & Privacy Committee
EM	Ethics Manager
GDPR	EU General Data Protection Regulation
EOSC	European Open Science Cloud
EU	European Union
EUPL	European Union Public License
GPL	General Public License
HER	Home energy reports
H2020	Horizon 2020
LGPL	Lesser General Public License
ORCiD	Open Researcher and Contributor ID
PID	Persistent identifier
PPC	Public Power Corporation S.A.
PSI	Public sector information
RCT	Randomized controlled trial
REC	Research Ethics Committee
SSH	Social Science and Humanities
URI	Uniform Resource Identifier

Executive Summary

This deliverable is an outcome of Task 5.2 "Data documentation" and aims to present the data collected and utilised by the EVIDENT project. It is the second of two deliverables, along with deliverable 5.3 "Data documentation", which is focused on presenting the relevant literature on data quality and key data quality indicators, the various data documentation approaches, the techniques for enhancing the discoverability and anonymisation of research data and provided a brief overview of the data created and collected by the EVIDENT project.

D5.4 'Updated data documentation' presents the main datasets of EVIDENT, their characteristics and how they will be made available to policymakers, researchers and the general public, while considering best practices for open science, legislative and operational requirements, the EVIDENT project grant agreement, etc. The deliverable also provides an overview of the open repositories used for sharing the data.

1. Introduction

Behaviour-based interventions for promoting energy efficiency among informed citizens can lead to improved social performance[1], client protection and, ultimately, greater environmental impact. Acknowledging these needs, the EVIDENT project is designed to further investigate the field, providing insights for optimized policy actions with significant societal impact.

One of the main goals of the EVIDENT project is to leverage data to support energy policies. In this context, the role of social science and humanities (SSH) should be taken into consideration. SSH studies can provide valuable insights into the social, economic, and cultural factors influencing energy consumption and renewable energy adoption. For example, SSH studies may involve analysing survey data to understand public attitudes towards renewable energy or examining policy documents to identify barriers to increasing energy efficiency.

The dissemination of SSH research and results can have an impact on policymaking. In this context, effective communication is crucial for European projects, especially for receiving policy recommendations from various audiences [1]. Open data can also assist in increasing transparency, accountability, and public participation in policy-related decisions. By making data openly available, it is possible to increase collaboration and innovation and help ensure that policy decisions are based on sound evidence.

Despite the prevalence of data value, there are still some questions regarding the quantification of data value and how it is evaluated. The answers could be given through the law of value and the efforts required to acquire or create the datasets [2], [3]. As discussed in D5.3 “Data Documentation”, it should be simple for people to discover and comprehend the available datasets in order for them to decide whether these could be useful prior to downloading them.

The value of data has a significant impact on the efficiency and performance of the various research processes. Therefore, data is usually considered a strategic resource, essential for organisations. In addition, the monetisation of data in the business or research ecosystem provides and generates considerable revenue.

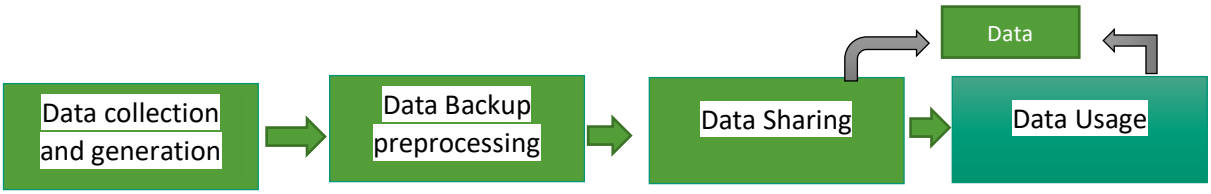


Figure 1: Data Sharing Value Chain [6]

[1] https://cordis.europa.eu/article/id/88588-better-business-strategy-for-improved-social-performance

The value of data can be understood by knowing the data-sharing value chain, as depicted in Figure 1. These activities transform the original raw data and can support innovation, but also increase the data value through usage and sharing. The value of data can be captured from the above activities, which transform the raw data into actions that lead to improved business processes and innovation. Single individuals may find it challenging to create value; therefore, they often need to interact with the entire value chain activities.

Data is also a significant aspect of research. Raw data can be shared through various platforms to increase their value by being leveraged by other researchers. For example, energy consumption data of public sector buildings can be shared, in order to be used by digital twin developers. In this case, the datasets of the applied transformations can be added to the original dataset, multiplying its value. Moreover, the public availability of the datasets can facilitate their validation by the scientific community. Furthermore, the foundations for creating communities around these datasets can be set.

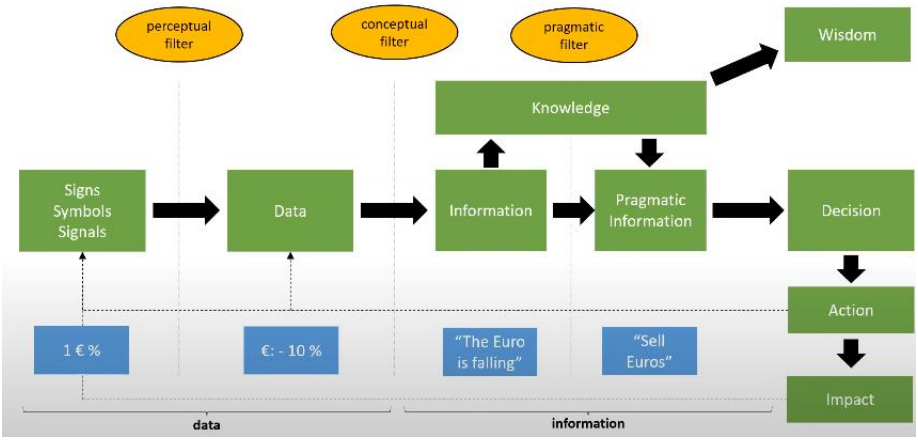


Figure 2: Transforming data into impact [6]

Figure 2 depicts the steps including the raw data generation and their transformation into pragmatic information that leads to informed decisions, actions, and impacts. Apart from having data available, it is important to know how to utilise them according to the particular use cases. For example, an autonomous machine could use this data for training. Also, a conceptual filter, which assigns the data to the relevant context, should be in place.

Similarly, the data should have features suited to the particular environment in which it will be utilised. Knowledge graphs are useful knowledge representation tools [4]. A knowledge graph is a plot of information that includes connections between related things, thereby facilitating their interpretation depending on the environment in which they are focused [3].

1.1 Data documentation for Evaluation policy

This deliverable is concerned with the documentation of the data that will be used for evaluating and assessing the policy interventions within the EVIDENT project. Transparency can be enhanced by making data understandable and openly accessible to the public. Open data access also motivates researchers to conduct additional analyses using the same data. Finally, particular attention is paid to ensuring the safety of sensitive information by anonymizing participant information before making their data openly accessible.

Nowadays, there is a consensus on the merits of using evidence for policymaking. In particular, using data to support policymaking can improve decision-making as decisions can be grounded in reality instead of assumptions or opinions. Additionally, data can be used to monitor policy outcomes, identify emerging trends, and increase accountability and transparency by providing information on how decisions are made.

In the European Union (EU), there is an ongoing effort to increase evidence-based policymaking. Such an example is the EU's Better Regulation agenda where one of the main goals is to 'ensure that policymaking is based on evidence[2]. In addition, the Horizon research and innovation funding programmes support research and innovation activities that can provide scientific evidence to inform EU policymaking. Other examples of funding programmes include the European Social Fund, which supports social policies based on evidence-based research and analysis, and the European Regional Development Fund, which supports evidence-based policies to promote sustainable regional development.

In the EVIDENT project, five use cases are developed that in effect summarise the project's proposed policy interventions. The EVIDENT use cases are:

- **Use Case 1 - Estimate the importance of consumption feedback in residential users:** The first use case of the EVIDENT project aims to estimate the importance of consumption feedback in residential users. The primary aim is to help households save energy and money through zero-cost interventions introducing mechanisms such as consumption feedback to provide information about historical data on energy consumption, and personalized energy conservation tips in a simplified manner. Mechanisms such as Home Energy Reports (HER) and digital platforms are employed to achieve the previous goals.
- **Use Case 2 - Estimate the relative effectiveness of interventions like peer comparison feedback**: The second EVIDENT use case aims to estimate the relative effectiveness of interventions like peer comparison feedback. The primary aim is to help households save energy and money through zero-cost innervations introducing mechanisms such as consumption feedback to provide information about historical data on energy consumption, and personalized energy conservation tips in a simplified manner. Mechanisms such as Home Energy Reports and digital platforms are employed to achieve the previous goals.
- **Use Case 3 - Investigate the role of big data in assessing the impact of behavioural insights in energy consumption**: The role of big data in assessing the impact of behavioural insights in energy consumption. Large amounts of energy production and consumption data are generated, collected and stored. This provides the possibility to implement energy big data mining and analysis. A set of socioeconomic, demographic and real-time energy consumption data can be combined to assess the impact of behavioural insights on energy consumption patterns. This high level of understanding of energy usage provides knowledge of good and secure investments for bridging the energy efficiency gap and means for verification of

[2] https://commission.europa.eu/law/law-making-process/planning-and-proposing-law/better-regulation_en

investment results. EVIDENT use case 3 builds on this approach by using machine learning models trying to reveal household energy consumption patterns. An additional approach followed in use case 3, enables the analysis of data coming from experimental settings (such as the field experiment in use case 1) and the examination of heterogenous treatment effects.

Use Case 4 - Examine the relation of energy consumption behavioural biases with consumers' financial literacy level: The fourth use case tries to estimate the relation of energy consumption behavioural biases with consumers' financial literacy level. Consumers often make energy investment choices that do not maximise their welfare, buying cheaper, less efficient appliances with higher running costs over time rather than more expensive, highly efficient appliances with lower running costs. This is referred to as the 'energy paradox' and describes the low adoption of energy-efficient technologies despite cost savings for consumers. Financial literacy, or the knowledge and skills to make effective economic decisions, is a key factor that is assumed to influence this energy efficiency gap. Another factor that may influence this energy efficiency gap is environmental literacy or environmental awareness, skills, knowledge, abilities and attitudes of consumers. The impact of economic and environmental literacy on household appliance decision-making remains poorly studied, particularly when consumers are faced with repair/replacement decisions. In addition, the analysis of the factors influencing repair/replacement decisions for rental properties remains understudied, even though 30% of European citizens live in such properties[3]. Using a serious game, this use case seeks to examine the impact of economic and environmental literacy on repair/replacement decisions for residential appliances. Through this, new intervention approaches and policies to enhance energy efficiency will be developed.

- **Use Case 5- Investigate and exploit energy demand curves**: Investing in energy-efficient household products, such as electrical appliances and small electronic products, often requires consumers to pay more upfront for more efficient products than for less efficient appliances, although overall cost savings are achieved over the lifetime of the product due to significantly lower operating costs. Therefore, when selecting a new appliance, the consumer must determine the total cost of an appliance over time to determine which is the most efficient. However, when selecting home appliances, consumers often do not choose the most advantageous option for them but rather the least costly option up front. This inefficient decision-making is believed to be due to discounting over time or the tendency to choose the cheapest option now over the cheapest option over time. This bias towards certain, immediate rewards as opposed to larger, uncertain rewards appears to be a key factor influencing the adoption of energy-efficient home technologies. An analysis of these discount rates and the factors that influence them is needed to increase the adoption of energy-efficient appliances. This use case seeks to investigate, through the use of a series of choice experiments, the impact of energy financial literacy, demographic factors and environmental literacy on discount rates and willingness to pay for efficient home appliances. The following methodology will be used.

[3] Data are from Eurostat (2021)

An overview of the use cases and the preliminary results can be found in D5.1 "Impact evaluation plan and policy measures", along with the monitoring of the project's key performance indicators. The use-case approach is designed to be straightforward and provide insights on real-world conditions, thus, can also inform the validity of EVIDENT's approach. For all cases, the effectiveness of the policy intervention is measured and reported.

In this context, sharing research data and efficient and effective data documentation is essential. For example, the Horizon 2020 programme requires that its projects emphasise dissemination and exploitation of results and promote open science. For the latter, data need to be deposited in trusted repositories which can provide free and open access with exceptions for certain types of data (sensitive, commercially obtained etc.). For example, some EVIDENT datasets are provided under the Creative Commons licence 'CC-BY', where the use of the datasets "… allows refusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use."[4]

Data documentation plays an important role in increasing datasets accountability, transparency, and usefulness. In particular, sharing research data enables other researchers to use and build on the data to create new knowledge. Also, sharing data promotes reproducibility, allowing the replication of the results using the same data and methods; thus, it allows independent validation of the findings. For this, Horizon 2020 beneficiaries are encouraged to follow the Findable, Accessible, Interoperable, Reusable (FAIR) principles, which aim to make scientific data more accessible, shareable, and reusable. The FAIR principles emphasize the importance of making research data available so that it can be easily discovered, understood, and used by humans and machines.

An overview of the FAIR principles is provided below [5]:

- **Findable**: Making research data findable by both humans and machines. Additionally, metadata describing the data, such as the author, title, and creation date, should be provided. The metadata should be machine-readable and standardized to enable efficient search and retrieval.
- **Accessible**: Making data openly accessible. Research data should be available to humans and machines, under clearly defined conditions of use, in an appropriate format for reuse.
- **Interoperable**: Making data interoperable. The data should be structured and formatted to enable integration with other datasets. Additionally, data should be available in a format that allows for automated processing, such as through application programming interfaces (APIs).
- **Reuse**: Increasing data reuse. The data should be able to be used for multiple purposes beyond the original research for which it was created.

[4] Source: https://creativecommons.org/about/cclicenses/

[5] For more see European Commission, Directorate-General for Research and Innovation, Turning FAIR into reality: final report and action plan from the European Commission expert group on FAIR data, Publications Office, 2018, https://data.europa.eu/doi/10.2777/1524

The FAIR principles have been adopted or are recommended by several organisations such as the European Open Science Cloud (EOSC), Environmental Systems Research Institute (ESRI), European University Association, Research Data Alliance (RDA), and are included in the Open Data Policy.[6]

1.2 Relation with other Deliverables and Tasks

This deliverable falls under Task 5.2 "Data documentation" and aims to present the data collected and utilised in the context of the EVIDENT project. It is the second of two deliverables and builds upon and expands Deliverable 5.3 "Data Documentation" which focuses on presenting the relevant literature on data documentation and key data quality indicators, the various data documentation approaches, the techniques for enhancing the discoverability of and anonymization of research data and provided a brief overview of the data created and collected by the EVIDENT project.

The deliverable receives as inputs D3.3 "Data collection and management", D5.1 "Impact evaluation plan and policy measures", and D5.3 "Data Documentation". In addition, the deliverable is closely related to all Work Package 5 "Policy measures" deliverables.

1.3 Structure of the Document

This deliverable is structured as follows:

- Section 2 - Data documentation. This section provides an overview of available open data repositories and open data and evaluation policies.
- Section 3 - Data structure. This section presents the main EVIDENT project datasets and their main characteristics. Additionally, it highlights the importance of the research results and data transformations into the raw data.
- Section 4 - Data access and sharing. This section focuses on the data sharing of the EVIDENT datasets, along with data ethics and privacy considerations.
- Section 5 – Conclusion. This section provides a brief overview of the deliverable.

[6] https://open-research-europe.ec.europa.eu/for-authors/data-guidelines#fairdata

2. Data Documentation

This section will describe the tools and platforms used for sharing several EVIDENT project datasets, including Zenodo (https://zenodo.org/) and other OpenAIRE repositories, to create the data documentation required for the evaluation [5]. These platforms provide transparency and re-usability of the data, in accordance with the FAIR principles. Additionally, this section includes a survey on research data platforms and data sharing through APIs and the evaluation policy.

The Zenodo platform can be used to share research results, including documents and datasets. Each item is assigned a Digital Object Identifier (DOI), which is a persistent identifier. In addition, there are other types of identifiers, such as researcher identifiers (e.g., ORCiD[7]). The above is important for the EVIDENT data generated by the platform, as they will be linked to ORCiD, and the participants will actually be able to get to know the research work of the organiser. Participants in the surveys from the platform will be able to review the ORCiD of the organizer and see the work he has done and to know the organizer better from the history of his work. In addition, it will be available and searchable both through the search engines as well as through the search on the platform. Besides EVIDENT data, we can use data generated by the platform. However, there are the EVIDENT data which are the data of the experiments of use case 3 the environmental data of Swedish cities and use cases 4 and 5. Furthermore, we will refer to the possibility offered by the platform to generate data for Zenodo.

Zenodo is compatible with DataCite, a metadata scheme that is standardized set of metadata fields describing research data such as the title, author, publication data, publisher, identifiers keywords and licence information. The DataCite metadata schema includes fields such as:

- **Title:** The title of the dataset;
- **Author:** The organization or the person who contributed the dataset;
- **Publication date:** The date when the dataset was published or made available;
- **Publisher:** The entity responsible for publishing or hosting the dataset;
- **Identifiers:** Persistent identifiers that identify the dataset, such as DOI or Uniform Resource Identifiers (URIs);
- **Keywords:** Descriptive terms or phrases that represent the main topics or themes of the data set;
- **License information:** The license or term of use for the dataset.

These fields have to be filled in in the Zenodo platform, during the dataset upload and can be used for facilitating the sharing. Also, EVIDENT adopts the Open Data, Software and Code Guidelines that are described as follows [6]:

1. The dataset(s) must be deposited in an appropriate data repository.
2. The dataset(s) must have a license applied which allows reuse by others (CC0 or CC-BY).
3. The dataset(s) must have a persistent identifier (e.g. a DOI), allocated by a data repository.

[7] ORCiD stands for Open Researcher and Contributor ID. It is a global, not-for-profit organization. “ORCID’s vision is a world where all who participate in research, scholarship, and innovation are uniquely identified and connected to their contributions across disciplines, borders, and time.” Source: https://info.orcid.org/what-is-orcid/

4. A Data availability statement must be included.
5. The dataset(s) citation reference must be included.
6. The dataset(s) should not contain sensitive information for the identification of the research's participants.
7. Related software and code should be also shared.
8. The dataset(s) must be useful and reusable by others, adhering to any relevant data-sharing standards and the FAIR principles.
9. The dataset(s) should link back to the research paper, if possible.

The aspects of long-term access and the use of open formats ([18]) for storing data are documented in D5.3, section 3.

2.1 Survey to Data Sharing Platform

This subsection provides an overview of widely used platforms that can enable researchers to upload and share their research outputs and datasets. Also, most of these platforms offer additional features to increase the discoverability of the uploaded data.

2.1.1 The European Open Science Cloud

The European Open Science Cloud (EOSC) (https://eosc-portal.eu) is a platform recognized by the Council of the European Union for uploading, storing, and processing research data that was designed to support open science in the EU. The platform adheres to the FAIR data principles and ensures seamless access and secure reuse of the uploaded data (e.g., source code, datasets, and publications). EOSC aims to establish an open multi-disciplinary environment where EU researchers, developers, as well as the general audience can publish, access, and reuse services, tools and data for education, research, and innovation purposes.

2.1.2 Zenodo

Zenodo is a widely used repository managed by CERN that enables researchers to upload and share their data and research outputs. It supports various kinds of data, such as images, source code, text, and large datasets. Data uploaded using Zenodo are stored in CERN's data canters. Furthermore, Zenodo promotes the discoverability and accessibility of shared data by assigning DOIs, when needed. A versioning feature is also supported, which facilitates uploading updated versions of the data. Moreover, Zenodo has integration functionalities with other platforms and tools, including ORCID and GitHub. Finally, the Zenodo platform is open source and is based on the Invenio open-source digital library tool. Zenodo leverages the flexibility and scalability of Invenio [7], [8] to provide a user-friendly interface and integration with other research tools and platforms to support various data formats and types. Invenio currently covers a suite of three core products developed by the Invenio community:

- InvenioRDM - a repository/document management platform
- InvenioILS - an integrated library system
- Invenio Framework - a code library for building large-scale information systems such as InvenioRDM and InvenioILS.

Besides the three primary outcomes, a more extensive number of less generic code libraries are also maintained for the benefit of the open-source community, independently of Invenio. Moreover, Inveio's capabilities can be applied to various policy areas, such as environmental sustainability,

economic development and social well-being. For example, by analysing satellite imagery, meteorological data and socio-economic indicators, Inveio can become a vehicle to provide information on the impacts of climate change, identify areas vulnerable to natural disasters and support the development of resilient infrastructure. With its ability to process and interpret complex data, Inveio enables policymakers to navigate the complexity of today's world and make informed decisions that lead to positive changes.

2.1.3 FigShare

FigShare (https://figshare.com) is an online repository that enables researchers to share their outputs in a discoverable and citable way. Various data are supported, including images, figures, source code, and datasets. Additionally, it provides integrations with other platforms such as GitHub and ORCID. Also, relevant research metrics are supported and provided by the Altmetric tool. Finally, an enhanced service, called FigShare+ is provided, that supports the upload and sharing of very large datasets at a one-time data publishing charge.

2.1.4 IEEE DataPort

IEEE DataPort (https://ieee-dataport.org) is an open-access data repository tailored to technology, computer science, and engineering researchers to upload, share, and find datasets. It also equips researchers with various tools for managing their research data, such as version control, metadata editing, and access control. Also, IEEE DataPort facilitates researchers to associate their data with scientific publications (e.g., standards, journals, and conferences), thereby promoting research transparency and reproducibility. Finally, IEEE DataPort supports DOI assignment, data curation, and long-term preservation towards ensuring high data quality and increasing accessibility and reusability.

2.1.5 Mendeley Data

Mendeley Data (https://data.mendeley.com) is an open community-based repository where researchers can store datasets, ensuring high accessibility. Various open licenses are supported, as well as the assignment of DOIs. Additionally, the datasets can be annotated with several metadata fields, such as titles, descriptions, administrative metadata, instructions etc. Mendeley Data is being maintained by Elsevier Inc, while the file/data repository is hosted on the Amazon Web Services platform.

2.1.6 Code Ocean

Code Ocean (https://codeocean.com) is a platform that enables researchers to collaborate more efficiently by maintaining comprehensive records of the source code and associated datasets. In addition, it offers computational capabilities; researchers can use these computational resources to run and reproduce the experiments in a cloud environment. Multiple programming languages, such as R, MATLAB, and Python are supported.

2.1.7 GitHub

GitHub (https://github.com) is a widely used web-based platform that provides collaboration tools and version control functions for software development. It enables developers to share source code repositories and track changes over time. While it is mainly used for source code storage, sharing, and management, it can also be used for storing and sharing datasets.

2.1.8 GoTriple Platform

The GoTriple Platform (https://www.gotriple.eu/) was developed by the French National Centre for Scientific Research and aims to serve as a reference platform for institutions, enterprises, media, and researchers. The platform includes tools and processes for supporting research (e.g., annotation tools, recommendation systems, visualisation techniques), discovering and reusing SSH outputs and resources (e.g., publications and datasets), and interacting with researchers from other disciplines.

2.2 Open Data and Evaluation for Policy Making

EU Directive No. 2003/98/EC on the reuse of public sector data is commonly known as the Public Sector Information (PSI) Directive and is often referred to as the 'open data directive' [9]. The Directive is an attempt to remove barriers to reusing data from the public sector. This Directive was amended in 2019 and has been in force since July 2019 (EU Directive 2019/1024) [10]. Public services create and collect huge amounts of valuable data every day. The Directive emphasizes the open availability of data and its reuse for both commercial and non-commercial purposes. In addition, regarding data reuse, Article 5 of the Directive encourages public services to create data in accordance with the principles of being open by design and by default. The Directive does not distinguish who can benefit from the data and does not discriminate or restrict markets and sections. Furthermore, the Directive mentions the idea of information and data as an infrastructure to promote innovation and transformation of open government data for reuse. The public sector should consider developing and investing in it, and their data assets should be leveraged and transformed into infrastructure resources.

Another important aspect is the definition of high-value datasets, as specified in Directive (EU) No. 2003/98/EC Annex I which includes the following:

- Geographical data
- Earth observation and environment
- Meteorological
- Statistics
- Companies and company ownership
- Mobility

A high-value dataset could include several sections listed above and plays an important role. Furthermore, the real-time availability of this data through APIs creates business opportunities, as indicated and promoted by the directive.

The use of open licenses is highlighted, combined with the fact that the public sector consists of various departments were making data available using different frameworks and approaches has created barriers to data reuse. The potential of open data is excellent, especially with the interconnection of heterogeneous data. However, the barriers are many, and we need a clear definition and Standardization.

The open data movement is based on the principle that data collected by public entities belong to the public and should therefore be made freely available to all. Open data aims to increase transparency, accountability and collaboration between government and citizens and promote economic growth and innovation. Open data allows citizens to gain insights into governed participation in decisions. Additionally, open data leads to higher efficiency and innovation in the private sector as businesses

can use public data to develop new products and services. Moreover, data can help researchers and academics to conduct more comprehensive and meaningful research. However, the data must be accurate and current, while sensitive information needs to be protected. Finally, the data should be accessible to everyone. It should be noted that any organization can make its data open, meaning that private data could be made available.

The goal of the evaluation policy is to ensure that policymakers' decisions are evidence-based, that policies are effective, efficient and achieve their intended outcomes. Open data and evaluation policy are closely related, focusing on promotional transparency, accountability and evidence-based decision-making. Making data freely available to the public open data can ensure that evaluations are based on accurate and up-to-date information. Open data can also promote transparency and accountability, as citizens can use public data to hold their elected officials accountable for their decisions. An evaluation policy can be used to assess the impact.

2.2.1 Open-source software

In the case of open-source software, the executable, as well as the source code, are shared, along with the appropriate documentation. Open-source software can also foster innovation by reducing the time to deliver value for strategic open-source projects, through co-development with industry customers for their specific use cases. A typical example is the Red Hat organisation, which owns and maintains a Linux distribution and provides the source code for that distribution. It is a notable example of how a commercial company benefits from open-source software. On one side, there exists an open community that has formed around the product and has helped develop it, while on the other side, the developers get the code and can use it under the license agreement that the company sets up. Additionally, communities are forming based on open-source repositories (e.g., Github), where developers can join and become contributors. They can create a copy of the official repository and retrieve its source code. Anyone can become a contributor, while some members are assigned the committer role, meaning that they review the code and decide what changes to make to the final version.

2.2.1.1 Software license

There are various open-source software licenses that can be adopted [11]. A license is a legal tool for managing the software considering its use, distribution, modification and reuse. Understanding the software license is vital for an organization, and it is mandatory for the organization to be aware of the licenses of the software it uses. Open-software licenses can be grouped into Copyleft and Permissive licenses. The Copyleft licenses enforce the same license on all products based on the source code. Permissive licenses like MIT, Apache, and Creative Commons enforce fewer restrictions. For example, a source code having a permissive license can be modified, while its license can be changed into a proprietary one. Furthermore, there exist licenses that cannot be changed, or enforce the open-source principles. Such licenses include General Public License, Mozilla Public License, and Lesser General Public License. These licenses prevent organisations from using the software without contributing back. A brief overview of some licenses is provided below:

GNU General Public License (GPL): This license enforces the freedom to execute, study, and modify software. Among other things, it includes the license to exploit it commercially. The updates to the GPL in Version 3.0 were instituted to protect the copyleft features given recent legal and technological changes.

GNU Lesser General Public License Version 3.0 (**LGPL-3.0**): LGPL is a weak copyleft license and can be applied to software deployed as a shared library [12]. The source code in the shared library must be available to be viewed, modified, and shared, but proprietary code using the library does need not be freely available.

Mozilla Public License Version 2.0 (**MPL-2.0**): A balance between proprietary and free software is provided by the MPL [12]. MPL-2.0 is a Copyleft license, similar to the GPL-3.0, facilitating the combination of source codes under different licenses. For example, software that is a mixture of proprietary and MPL code must only be distributed with changes to MPL-licensed files.

MIT License: The MIT license is a permissive free software license that originated at the Massachusetts Institute of Technology (MIT) [13], [14]. As a permissive license, it imposes limited restrictions on reuse and therefore has a high degree of license compatibility [9]. Unlike copyleft software licenses, the MIT license also permits reuse within proprietary software, provided that all software copies or substantial portions contain a copy of the MIT license terms and a copyright notice. According to [15] and [16], it is considered the most used software license in the GitHub repository.

European Union Public License (EUPL): Software developed by the European Commission is distributed under this license [17]. The purpose of the EUPL is not to compete with any of these licenses, but to encourage a new generation of public administrations to adopt the Free/Open Source model to exploit their software and knowledge, starting with the European institutions directly. A typical example is the action plan for smart cities and regions for 2023-2026 from Slovakia, which promotes the use of open-source software such as the EUPL [18]. By promoting an integrated and innovative environment, the plan enhances the capacity of public sector contractors, facilitates knowledge transfer and supports the development of smart cities and regions.

3. Data Structure

This section presents the EVIDENT datasets and outlines the data structure, as well as the transformation of the data into valuable insights. The structure of the results has a major role, especially when the results are to be shared. Having the raw data and applying transformations will make the transformed data searchable and valuable, as well as the transformations that have been applied to be evaluated and annotated by the community that uses them. This procedure is useful when both datasets are shared and will be used in multiple cases. First, we have to define and manage the quality of the data and understand the information production process. Figure 3 shows the process of transforming raw data into high-valued information in a healthcare scenario.

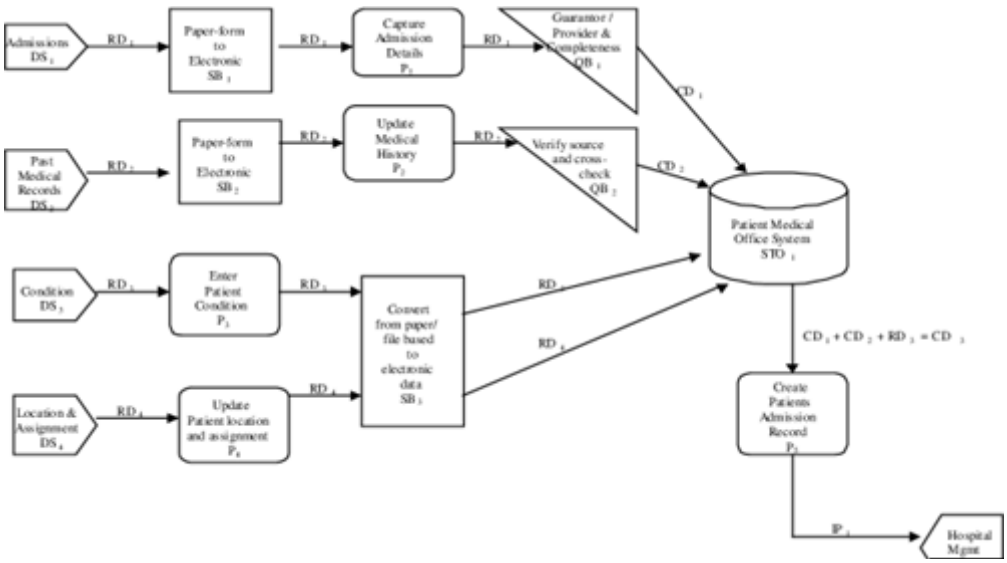


Figure 3: Raw data into high-value information [19]

The final outcome of this process is the patient record. Data can be created by humans or machines (e.g., smartphones, cars, sensors). The use of data, which occurs when searching and accessing them, creates an added value. In addition, different datasets can be aggregated. For example, data from buses or trains can be retrieved and combined with weather or traffic data to estimate potential delays. This integration of data from different domains can lead to high-value insights.

The EVIDENT project, through its 5 large-scale use cases, seeks to evaluate the impact of behavioural insights in energy efficiency and conservation. Using different methodologies and tools, such as Randomized Control Trials (RCTs), quasi-experiments, serious games and bid data analytics, the project consortium tries to estimate the causal impact of behavioural biases on energy efficiency. For the design and implementation of all EVIDENT use cases, three main sources of data were leveraged: (a) the two energy companies[8], (b) the EVIDENT platform[9] developed to cover the data collection

[8] Detailed information about the design, the implementation, the analysis and the results of all use cases can be found in the deliverables of WP2, WP3 and WP4 here.

[9] Technical information about the EVIDEN platform can be found in all deliverables of WP6 here. The EVIDENT platform can be found at the following address: https://platform.evident-h2020.eu

needs of the project, and (c) third-party services. The rest of the section presents the datasets collected and used from the EVIDENT partners for each use case.

3.1 EVIDENT Dataset for use case 1 and use case 2

Use cases 1 and 2 aim to estimate the importance of consumer feedback and peer comparison feedback in household energy consumption. The main objectives of both use cases is to explore the effectiveness of nudges in energy conservation, raise consumers' awareness regarding energy consumption, and provide energy consumption tips and suggestions for more efficient use of electricity. To do so, the EVIDENT partners implement an RCT in Sweden and analysis on the available data provided by two project partners, CW and PPC. CW provides data for around 900 Swedish prosumers[10], including energy-related measurements, such as energy production and consumption, as well as a set of demographic data for these prosumers. On the other hand, PPC provides energy consumption measurements for its customers and a rich set of demographics from the "myEnergyCoach" platform. Tables 1 and 2 present CW's datasets, while Table 3 and Table 4 describe the data provided by PPC.

Table 1: CW Prosumers Energy Measurements Dataset

Dataset Name	**CW Prosumers Energy Measurements Dataset**
Dataset Summary	
Dataset Description	This dataset includes anonymized hourly energy measurements for around 900 Swedish prosumers. Each record of the dataset includes an anonymized client identifier, the timestamp for the referred hour (e.g., 20/04/2023 15:00), the energy bought, produced and sold in watthours. Finally, it includes the total energy consumption which is calculated as the following: $$\text{Total consumption} = \text{Bough} + \text{Produced} - \text{Sold}$$
Dataset Type/Format	Alpha-numeric values saved in a MySQL database
Dataset Origin	Private databases and repositories of CW
Expected Size	The dataset consists of 36million records (1.6 GB)
FAIR Data - Making data findable, including provisions	
Discoverable and Identifiable Data	-
Naming Convention	CW_MEASUREMENTS
Versioning	The database will continuously be updated till the end of the project
Search Keywords	-
Metadata	-
Geographical Scope	Sweden
Time period	Energy measurements from 01/01/2000 to 06/04/2023. Note: Data before December 2020 is not used in the EVIDENT project.

[10] Prosumer: An individual who both consumers and produces electricity.

FAIR Data - Making data openly accessible	
Dataset Openly Accessible	The dataset will not be openly accessible.
Access Restrictions	Access to the dataset is limited to partners that participate in the pilots of CW and conduct the analysis, on a need-to-know basis.
Repository	-
FAIR Data – Increase data re-use	
Licence	Private license based on bilateral agreement
Data sharing terms	Private license based on bilateral agreement
Re-use timeframe	The database will continuously be updated till the end of the project
Data Security	
Security Measures	The MySQL database the data are safely stored is accessible only through a virtual private network and after authentication.
Privacy Measure	Anonymization is applied to clients' identifiers. The mapping between the anonymised identifiers and the customers is only known to the corresponding department of CW.
Ethical Aspects	
Ethical and Legal Consideration	-
Consent form for sharing the dataset publicly.	-
Other Issues	-

Table 2: CW Prosumers Demographics Dataset

Dataset Name	**CW Prosumers Demographics Dataset**
Dataset Summary	
Dataset Description	This dataset includes the corresponding demographic data and information about the building size, household members, heating system types for the 900 prosumers presented in Table 1.
Dataset Type/Format	Alpha-numeric values saved in a MySQL database
Dataset Origin	Private databases and repositories of CW
Expected Size	The dataset consists of 900 records (0.17 GB)
FAIR Data - Making data findable, including provisions	
Discoverable and Identifiable Data	-
Naming Convention	CW_CONSUMERS
Versioning	The database will continuously be updated till the end of the project
Search Keywords	-
Metadata	-
Geographical Scope	Sweden
Time period	These data do not include the dimension of time.
FAIR Data - Making data openly accessible	

Dataset Openly Accessible	The dataset will not be openly accessible.
Access Restrictions	Access to the dataset is limited to partners that participate in the pilots of CW and conduct the analysis, on a need-to-know basis.
Repository	-
FAIR Data – Increase data re-use	
Licence	Private license based on bilateral agreement
Data sharing terms	Private license based on bilateral agreement
Re-use timeframe	The database will continuously be updated till the end of the project
Data Security	
Security Measures	The MySQL database the data are safely stored is accessible only through a virtual private network and after authentication.
Privacy Measure	Anonymisation is applied to clients' identifiers. The mapping between the anonymised identifiers and the customers is only known to the corresponding department of CW.
Ethical Aspects	
Ethical and Legal Consideration	-
Consent form for sharing the dataset publicly.	-
Other Issues	-

The energy measurements, provided by CW, are collected automatically through the smart meters supplied to their clients. The smart meters send information regarding the produced energy from the photovoltaic installations and the amount of the energy transferred to the grid (energy produced but not consumed by the prosumers). Information about the energy bought by each prosumer is provided by Swedish Distribution System Operators (DSOs) to CW. This information is used so the total consumption is calculated.

The data are provided in hourly format, so the EVIDENT partners aggregate them into daily and weekly observations to satisfy the needs of the experiments. While data collection is an automated process, it is not unlikely that the smart meters fail to deliver valid data either due to hardware malfunctions or software issues. Thus, the data collected might not be accurate. In that case, the whole record will be dropped, and the corresponding values will be replaced by the average of the previous and the next observations. This is also the case for missing values. Additional information about the Dataset can be found in Appendix 0 "CW Dataset".

Table 3: PPC Consumers Energy Measurements Dataset

Dataset Name	**PPC Consumers Energy Measurements Dataset**
Dataset Summary	
Dataset Description	The dataset includes 25000 entries of energy consumption values from Greek households where each record represents a unique consumer. Additionally, there is information from each household regarding the consumption of each device. The consumption is measured in kWh (kiloWatthours)

Dataset Type/Format	Numeric (integer, float) values stored in a CSV database
Dataset Origin	Private database of PPC
Expected Size	Approximately 350MB
FAIR Data - Making data findable, including provisions	
Discoverable and Identifiable Data	-
Naming Convention	MyEnergyCoach_consumption_data_export
Versioning	The database is steadily growing and will be updated until the end of the project, as more and more consumers will join the platform. (version_001)
Search Keywords	consumption, meter id, device id, device consumption
Metadata	No metadata
Geographical Scope	Greece
Time period	The platform is running from early 2020 until present. All the data will be used for exploitation and analysis for the project's scope
FAIR Data - Making data openly accessible	
Dataset Openly Accessible	The dataset is private and will not be publicly accessible. An alternative version will be available in public in the future
Access Restrictions	The dataset is accessible and open to the consortium partners who participate in the project on behalf of PPC and work either for data collection or data analysis
Repository	Private PPC repository
FAIR Data – Increase data re-use	
Licence	Private license based on bilateral agreement
Data sharing terms	Private license based on bilateral agreement
Re-use timeframe	The database is constantly being updated until the end of the project
Data Security	
Security Measures	The database is safely stored and is accessible only by PPC members through VPN and dedicated password-encrypted servers
Privacy Measure	Anonymization protocol is applied to all records. The proper mapping between the meter identification numbers and the consumers is known to the department responsible for the data analytics
Ethical Aspects	
Ethical and Legal Consideration	-
Consent form for sharing the dataset publicly.	-
Other Issues	-

Table 4: PPC Consumers Demographics Dataset

	PPC Consumers Demographics Dataset
Dataset Summary	
Dataset Description	The dataset includes all data obtained through an online demographic template that stores total consumption, household characteristics, and information about the housekeepers. The proposed dataset offers a wide range of household demographic information, including construction information such as floor, square meters, built year, windows panel type, appliance data like hot water system, heating system, lights frequency, a list of electrical devices, and their usage frequency. It also includes information such as age group, the age of the person who submitted the information to the proposed template, the number of people living in each household, etc. (Figure 1)
Dataset Type/Format	Categorical format options through predefined (multiple) choice answers stored in a CSV database
Dataset Origin	Private database of PPC
Expected Size	Approximately 555MB.
FAIR Data - Making data findable, including provisions	
Discoverable and Identifiable Data	-
Naming Convention	MyEnergyCoach_demographics_data_export
Versioning	The database is steadily growing and will be updated until the end of the project, as more and more consumers join the platform and complete the forms.
Search Keywords	appliances, lights, answers, heating system, meter id, frames, square meters
Metadata	No metadata
Geographical Scope	Greece
Time period	The platform is running from early 2020 until present. All the data will be used for exploitation and analysis for the project's scope
FAIR Data - Making data openly accessible	
Dataset Openly Accessible	The dataset is private and will not be publicly accessible. An alternative version will be available in public in the future
Access Restrictions	The dataset is accessible and open to the consortium partners who participate in the project on behalf of PPC and work either for data collection or data analysis
Repository	Private PPC repository
FAIR Data – Increase data re-use	
Licence	Private license based on bilateral agreement
Data sharing terms	Private license based on bilateral agreement
Re-use timeframe	The database is constantly being updated until the end of the project
Data Security	
Security Measures	The database is safely stored and is accessible only by PPC members through VPN and dedicated password-encrypted servers
Privacy Measure	Anonymization protocol is applied to all records. The proper mapping between the demographic answers and the consumers is known to the department responsible for the data analytics
Ethical Aspects	

Ethical and Legal Consideration	-
Consent form for sharing the dataset publicly.	-
Other Issues	-

The dataset consists of about 25000 consumers with various characteristics that refer to a four-month total consumption per household. Data collection is an automated procedure. To provide feedback/recommendations about energy consumption, and achieve energy efficiency, a pre-processing step to enhance data quality is vital. The pre-processing step should deal with the dataset's defect that reduces the data quality and degrades the tool's performance in general.

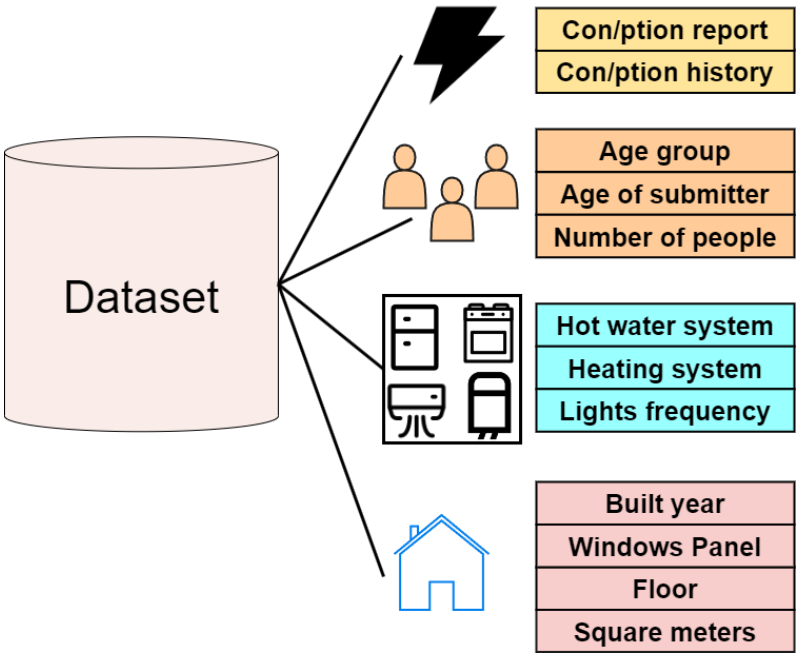


Figure 4: Dataset architecture

Although the dataset provides adequate data, some records of demographics and consumptions are not fully filled or filled with nulls or wrong format (e.g., text instead of a number, swapped answers in questions) as shown in Figure 4. To resolve these issues, each demographic category is analysed and only the entries with all fields properly completed are kept (data cleansing). Then, a section with the entries (meter ids) with all fields of consumption characteristics filled, comprises the final dataset. This process yields a smaller-sized dataset with 3000 out of 24980 households archived, as shown in Figure 5. Additional information about the Dataset can be found in Appendix 0 "PPC Dataset".

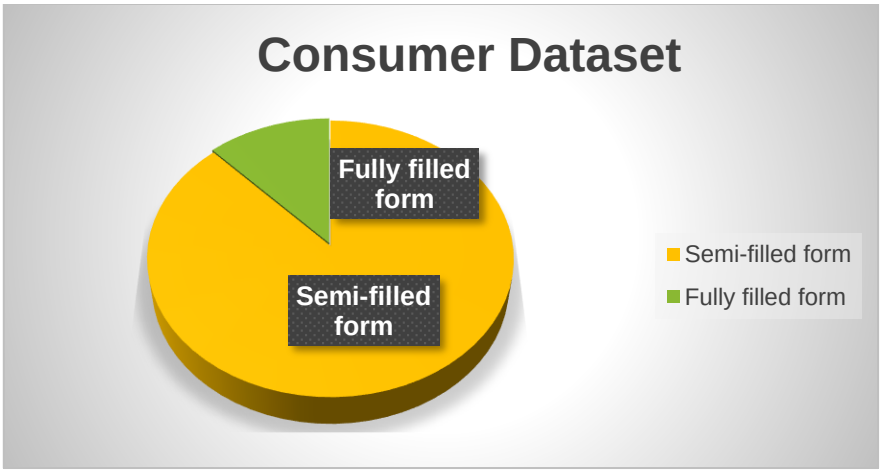


Figure 5: Percentage of fully/semi-filled forms

3.2 EVIDENT Dataset for Use Case 3

Use case 3 leverages artificial intelligence and machine learning methodologies to reveal hidden consumption patterns in the available data. In this use case, two analytical frameworks are developed to a) forecast energy consumption at a household level and b) identify the main drivers that affect consumers' performance in terms of energy conservation. CW provides the data in this use case and refers to energy consumption, production measurements, and available demographic data. In addition, weather data (weather condition data) collected by third-party services are used as extra control variables for the analyses of use cases 1, 2, and 3. Since the data provided by CW has already been presented in section 2.1, in this section we present only the weather data collected and used.

Table 5: Weather conditions for Sweden Cities

Dataset Name	Weather conditions for Sweden Cities
Dataset Summary	**Environmental Data from Swedish cities**
Dataset Description	Environmental data from 615 cities in Sweden includes 4000000 entries. For each city we collected the temperature, minimum temperature, pressure, humidity, wind speed, sunrise, sunset, and weather description. Time is not fixed; 6 to 5 measurements daily at random intervals.
Dataset Type/Format	CSV format
Dataset Origin	Data were created with an agent from an external API and belong to CERTH.
Expected Size	450MB
FAIR Data - Making data findable, including provisions	
Discoverable and Identifiable Data	Metadata tags, EVIDENT Community
Naming Convention	Sweden_ cities environmental
Versioning	05
Search Keywords	weather Sweden environmental data Swedish climate data Sweden weather data Sweden temperature data Sweden humidity data

	Swedish meteorological data Sweden city climate data Swedish city weather data Sweden city temperature data Sweden city humidity data Swedish urban climate data Sweden urban weather data Sweden urban temperature data Sweden urban humidity data Dataset of Sweden city temperatures and humidities Dataset of environmental data for 700 Swedish cities Swedish city climate and weather dataset
Metadata	Following the FAIR Principles, Zenodo uses a rich set of metadata so the dataset to be findable, accessible, interoperable and reusable. More information about Zenodo metadata can be found here.
Geographical Scope	Sweden
Time period	6 measurements per day
FAIR Data - Making data openly accessible	
Dataset Openly Accessible	Creative Commons
Access Restrictions	-
Repository	Zenodo link
FAIR Data – Increase data re-use	
Licence	Data sharing terms related to Creative Commons Attribution 4.0 International license. More specifically: 1. Attribution: The licensor must be credited in any reasonable manner, including by providing a link to the license and indicating any changes made to the licensed material. 2. Adaptation: The licensed material can be adapted or modified, as long as the resulting work is distributed under the same license. 3. Commercial use: The licensed material can be used for commercial purposes. No additional restrictions: No additional restrictions can be added to the license that would prevent others from exercising the permissions granted by the license.
Data sharing terms	Creative common license
Re-use timeframe	-
Data Security	
Security Measures	No Personal information
Privacy Measure	Does not need in this dataset there is no privacy issue
Ethical Aspects	
Ethical and Legal Consideration	None
Consent form for sharing the dataset publicly.	None
Other Issues	-

The data was captured by a custom-developed software agent that operates continuously and gathers data from an Application Programming Interface (API) for the Swedish cities. The agent stores the result in a PostgreSQL database. More information can be found it Appendix 0 “Environmental Dataset for Use Case 3”. Of note, due to technical issues, the data timeliness is not stable and some measures were lost during the day. However, the end result is not affected as there exists a considerable number of measurements in the day for each city. Furthermore, D5.3 “Data Documentation” and the DQMIS [20] were followed to create a variant length dimension for data quality like accuracy, reliability, timeliness, completeness, precision, and integrity. In the first version of this dataset named “swedish_cities_environmental.csv”, the quality was not high; however, it was corrected by removing the inappropriate rows to have more accuracy in the column of timestamp. Figure 6 depicts the respective Zenodo page for the dataset.

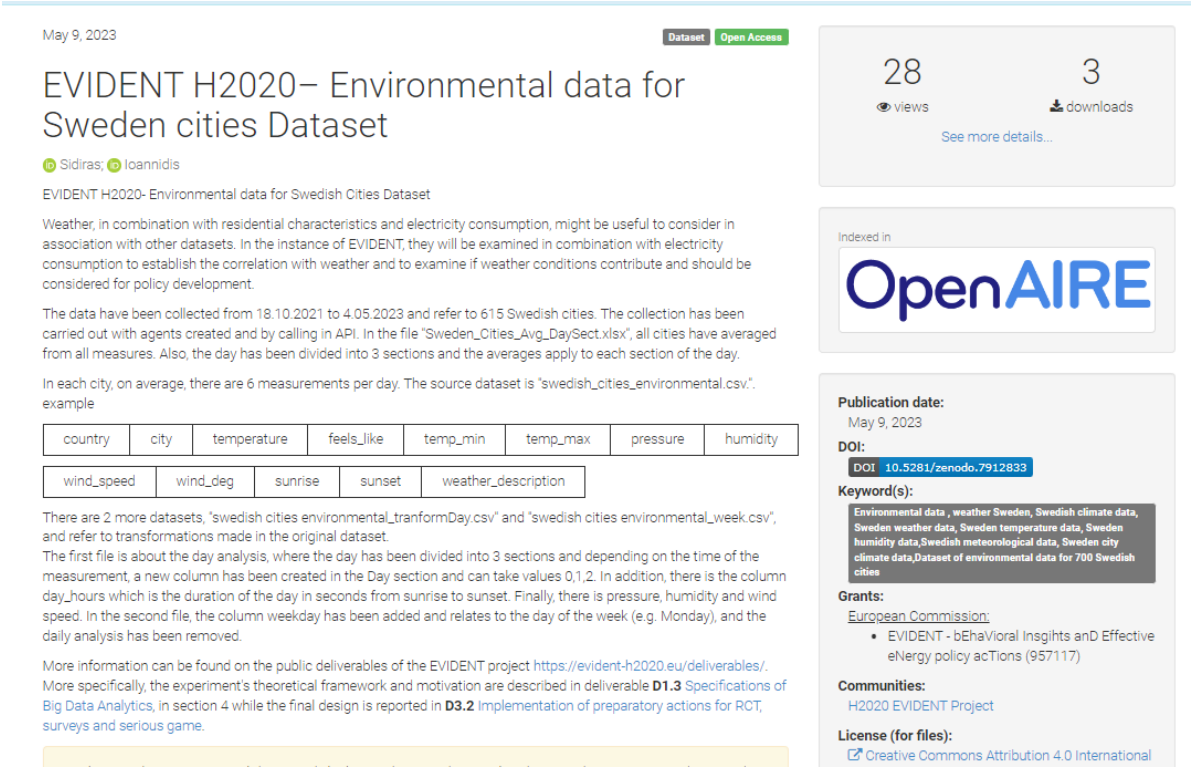


Figure 6: EVIDENT environmental data on Zenodo

3.3 EVIDENT Dataset for Use Case 4

Use case 4 aims to explore the impact of socio-demographic factors, environmental literacy, and financial literacy on consumer willingness to pay for the repair of home appliances. Through a serious game designed by the EVIDENT consortium and leveraging the EVIDENT platform's data collection services, the project’s consortium will collect and analyse a large sample of around 1000 respondents. Table 6 presents the data collected in the context of use case 4.

Table 6: Consumer Appliance Replacement Game

Dataset Name	Consumer Appliance Replacement Game
Dataset Summary	
Dataset Description	This dataset includes the data collected in the context of the EVIDENT Consumer Appliance Replacement Game. The experiment consists of the following sections: 1) demographic information; 2) financial literacy; 3) environmental literacy; 4) and the EVIDENT Serious Game.
Dataset Type/Format	Alpha-numeric values saved in an xlsx file
Dataset Origin	Data collected through the EVIDENT platform
Expected Size	The initial version of the dataset includes only 6 replies (15.4 kB), however till the end of the project, the dataset will be updated with all replies that will be collected (around 0.025 GB).
FAIR Data - Making data findable, including provisions	
Discoverable and Identifiable Data	DOI: https://doi.org/10.5281/zenodo.7956164
Naming Convention	CARG_REPLIES
Versioning	Version 1 (May 22, 2023). The dataset is expected to be updated till the end of the project.
Search Keywords	behavioural biases, environmental literacy, financial literacy, choice experiment, behavioural science, serious game, life-simulation game, energy efficiency, home appliances, repair or replace, energy consumption
Metadata	Following the FAIR Principles, Zenodo uses a rich set of metadata so the dataset to be findable, accessible, interoperable and reusable. More information about Zenodo metadata can be found here.
Geographical Scope	Several countries represented (among others, Sweden, Greece, Ireland, UK, Italy, France, Germany, USA)
Time period	The data collection process started on April 2023 and will last till September 2023.
FAIR Data - Making data openly accessible	
Dataset Openly Accessible	The whole dataset will be public available by the end of the project. Dataset's DOI can be used to locate the online repository where the dataset is hosted.
Access Restrictions	There are no access restrictions, however an attribution is recommended as "Delemere, Emma, Liston, Paul, Karypidis, Paris-Alexandros, & Pragidis, Ioannis. (2023). EVIDENT H2020 – EVIDENT Serious Games Dataset (0.1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7956164"
Repository	Zenodo: https://doi.org/10.5281/zenodo.7956164
FAIR Data – Increase data re-use	
Licence	Creative Commons Attribution 4.0 International (CC BY 4.0)
Data sharing terms	Data sharing terms related to Creative Commons Attribution 4.0 International license. More specifically: 1. Attribution: The licensor must be credited in any reasonable manner, including by providing a link to the license and indicating any changes made to the licensed material. 2. Adaptation: The licensed material can be adapted or modified, as long as the resulting work is distributed under the same license. 3. Commercial use: The licensed material can be used for commercial purposes.

	4. No additional restrictions: No additional restrictions can be added to the license that would prevent others from exercising the permissions granted by the license.
Re-use timeframe	The data will remain published in Zenodo repository.
Data Security	
Security Measures	The data are collected through the EVIDENT platform and are safely stored platform's database. The EVIDENT platform considers security by design.
Privacy Measure	The participation in the EVIDENT experiments is anonymous, thus no personal information or identifying information are collected through the platforms. Survey responses will be combined, and results will be reported as a group to ensure anonymity. In the public versions of this dataset, all personal and identifying information have been redacted.
Ethical Aspects	
Ethical and Legal Consideration	-
Consent form for sharing the dataset publicly.	-
Other Issues	Each participant should consent to a list of terms and conditions before participating to the experiment. Terms and conditions describe in details the purpose of the experiment, instructions on how to participate, who is running the experiment, participant's rights and data management.

The data collection process started in April 2022, and it is expected to be finished by September 2023. The analytical tools for the analysis of the collected data will be developed through the upcoming period and the final results will be reported in D4.3 "Updated econometric methodologies and robustness tests". In this experiment, consumers can participate in the experiment through the EVIDENT platform. After the collection of the data, the consortium members perform minor corrections in the responded values while most of the answers were provided to the participants as a list of options to avoid typos. Finally, all questions were mandatory for the participants so to avoid replies with missing entries. Figure 7 depicts the consumer appliance replacement game dataset Zenodo page.

Figure 7: Consumer Appliance Replacement Game on Zenodo repository

3.4 EVIDENT Dataset for Use Case 5

Use case 5 aims to identify the factors affecting consumers' willingness to pay for more expensive but efficient household appliances. These factors may relate to consumers' demographics and energy-related financial and environmental literacy levels. In the context of use case 5, two different e-lab experiments are designed and implemented, a) a Discrete Choice Experiment (DCE) in which each participant is presented with several hypothetical energy appliance choices (ovens, fridges, washing machines etc.), and is asked to identify the offer that they would choose, and b) a choice experiment to explore consumers Average Price Bias (APB). Table 7 and Table 8 present the two datasets collected for the DCE and the APB quasi-experiments in the context of use case 5, respectively.

Table 7: Discrete choice experiment dataset

Dataset Name	DCE Dataset
Dataset Summary	
Dataset Description	This dataset includes the data collected in the context of the EVIDENT Discrete Choice Experiment. The experiment consists of the following sections: 1) demographic information; 2) current home appliance purchasing behaviour; 3) financial literacy; 4) environmental literacy; 5) stated preference experiment consisting of four choice points; 6) discount rates; 7) discrete choice experiment consisting of ten choice points; and 8) questions examining direct rebound rates associated with the novel appliance selected.
Dataset Type/Format	Alpha-numeric values saved in an xlsx file
Dataset Origin	Data collected through the EVIDENT and the Qualtrics platform
Expected Size	The initial version of the dataset includes only 10 replies (17.8 KB), however till the end of the project, the dataset will be updated with all replies that will be collected (around 0.35 GB).
FAIR Data - Making data findable, including provisions	
Discoverable and Identifiable Data	DOI: https://doi.org/10.5281/zenodo.7825986
Naming Convention	DCE_REPLIES
Versioning	Version 1 (Apr 13, 2023). The dataset is expected to be updated till the end of the project.
Search Keywords	Behavioural Biases, Environmental Literacy, Financial Literacy, Choice Experiment, Behavioural Science
Metadata	Following the FAIR Principles, Zenodo uses a rich set of metadata so the dataset to be findable, accessible, interoperable and reusable. More information about Zenodo metadata can be found here.
Geographical Scope	40 countries represented (among others, Sweden, Greece, Ireland, UK, Italy, France, Germany, USA)
Time period	The data collection process started on November 2022 and will last till September 2023.
FAIR Data - Making data openly accessible	
Dataset Openly Accessible	The whole dataset will be public available by the end of the project. Dataset's DOI can be used to locate the online repository where the dataset is hosted.
Access Restrictions	There are no access restrictions, however an attribution is recommended as "Delemere, Emma, & Liston, Paul. (2023). EVIDENT H2020 – Discrete Choice Experiment Dataset [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.7825986"
Repository	Zenodo: https://doi.org/10.5281/zenodo.7825986
FAIR Data – Increase data re-use	
Licence	Creative Commons Attribution 4.0 International (CC BY 4.0)
Data sharing terms	Data sharing terms related to Creative Commons Attribution 4.0 International license. More specifically: 1. Attribution: The licensor must be credited in any reasonable manner, including by providing a link to the license and indicating any changes made to the licensed material.

	2. Adaptation: The licensed material can be adapted or modified, as long as the resulting work is distributed under the same license. 3. Commercial use: The licensed material can be used for commercial purposes. 4. No additional restrictions: No additional restrictions can be added to the license that would prevent others from exercising the permissions granted by the license.
Re-use timeframe	The data will remain published in Zenodo repository.
Data Security	
Security Measures	The data are collected through the EVIDENT and the Qualtrics platform and are safely stored in the corresponding databases. Both platforms consider security by design.
Privacy Measure	The participation in the EVIDENT experiments is anonymous, thus no personal information or identifying information are collected through the platforms. Survey responses will be combined, and results will be reported as a group to ensure anonymity. In the public versions of this dataset, all personal and identifying information have been redacted.
Ethical Aspects	
Ethical and Legal Consideration	-
Consent form for sharing the dataset publicly.	-
Other Issues	Each participant should consent to a list of terms and conditions before participating to the experiment. Terms and conditions describe in details the purpose of the experiment, instructions on how to participate, who is running the experiment, participant's rights and data management.

Table 8: Average price bias dataset

Dataset Name	**APB Dataset**
Dataset Summary	
Dataset Description	This dataset includes the data collected in the context of the EVIDENT Average Price Bias experiment. The experiment consists of five discrete key sections: 1) a section about participant's demographic data, 2) a small set of questions related to behavioural biases, 3) a set of questions related to financial literacy, 4) a section with questions related to environmental literacy and 5) the choice experiment about price perceptions.
Dataset Type/Format	Alpha-numeric values saved in an xlsx file
Dataset Origin	Data collected through the EVIDENT and the Qualtrics platform
Expected Size	The initial version of the dataset includes only 8 replies (28.5 KB), however till the end of the project, the dataset will be updated with all replies that will be collected (around 0.10 GB).
FAIR Data - Making data findable, including provisions	
Discoverable and Identifiable Data	DOI: https://doi.org/10.5281/zenodo.7825632
Naming Convention	APB_REPLIES

Versioning	Version 0.1.0 (Apr 13, 2023). The dataset is expected to be updated till the end of the project.
Search Keywords	quasi-experiment, average price bias, consumers' misperception, marginal prices, behavioural biases, financial literacy, environmental literacy, decision-making, nonlinear energy pricing, policy recommendations, nonlinear pricing schemes
Metadata	Following the FAIR Principles, Zenodo uses a rich set of metadata so the dataset to be findable, accessible, interoperable and reusable. More information about Zenodo metadata can be found here.
Geographical Scope	40 countries represented (among others, Sweden, Greece, Ireland, UK, Italy, France, Germany, USA)
Time period	The data collection process started on November 2022 and will last till September 2023.
FAIR Data - Making data openly accessible	
Dataset Openly Accessible	The whole dataset will be public available by the end of the project. Dataset's DOI can be used to locate the online repository where the dataset is hosted.
Access Restrictions	There are no access restrictions, however an attribution is recommended as "Pragidis, Ioannis, & Karypidis, Paris-Alexandros. (2023). EVIDENT H2020 - Average Price Bias Dataset (0.1.0) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.7825632"
Repository	Zenodo: https://doi.org/10.5281/zenodo.7825632
FAIR Data – Increase data re-use	
Licence	Creative Commons Attribution 4.0 International (CC BY 4.0)
Data sharing terms	Data sharing terms related to Creative Commons Attribution 4.0 International license. More specifically: 1. Attribution: The licensor must be credited in any reasonable manner, including by providing a link to the license and indicating any changes made to the licensed material. 2. Adaptation: The licensed material can be adapted or modified, as long as the resulting work is distributed under the same license. 3. Commercial use: The licensed material can be used for commercial purposes. 4. No additional restrictions: No additional restrictions can be added to the license that would prevent others from exercising the permissions granted by the license.
Re-use timeframe	The data will remain published in Zenodo repository.
Data Security	
Security Measures	The data are collected through the EVIDENT and the Qualtrics platform and are safely stored in the corresponding databases. Both platforms consider security by design.
Privacy Measure	The participation in the EVIDENT experiments is anonymous, thus no personal information or identifying information are collected through the platforms. Survey responses will be combined, and results will be reported as a group to ensure anonymity. In the public versions of this dataset, all personal and identifying information have been redacted.
Ethical Aspects	
Ethical and Legal Consideration	-

Consent form for sharing the dataset publicly.	-
Other Issues	Each participant should consent to a list of terms and conditions before participating in the experiment. Terms and conditions describe in detail the purpose of the experiment, instructions on how to participate, who is running the experiment, participant's rights and data management.

The data collection process for both datasets started in November 2022 and is expected to be finished in September 2023. The analytical tools for both experiments will be developed through the upcoming period and the final will be reported in D4.3 "Updated econometric methodologies and robustness tests".

The participants can participate in the experiment through two different platforms, the EVIDENT platform and the Qualtrics platform. Thus, a mapping mechanism between the two platforms was developed to unify the replies into a single format (a unique identifier was used in all questions and answers). After the collection of the data, following a manual process, the consortium members managed to gather the data into a single file and minor corrections in the responded values were implemented (e.g., "I don't know" and "I do not know"). Most of the answers were provided to the participants as a list of options to avoid typos and homogenize the responses. Finally, all questions were mandatory for the participants to avoid replies with missing entries. Figure 8 and Figure 9 depict the Zenodo pages for the APB and DCE datasets, respectively.

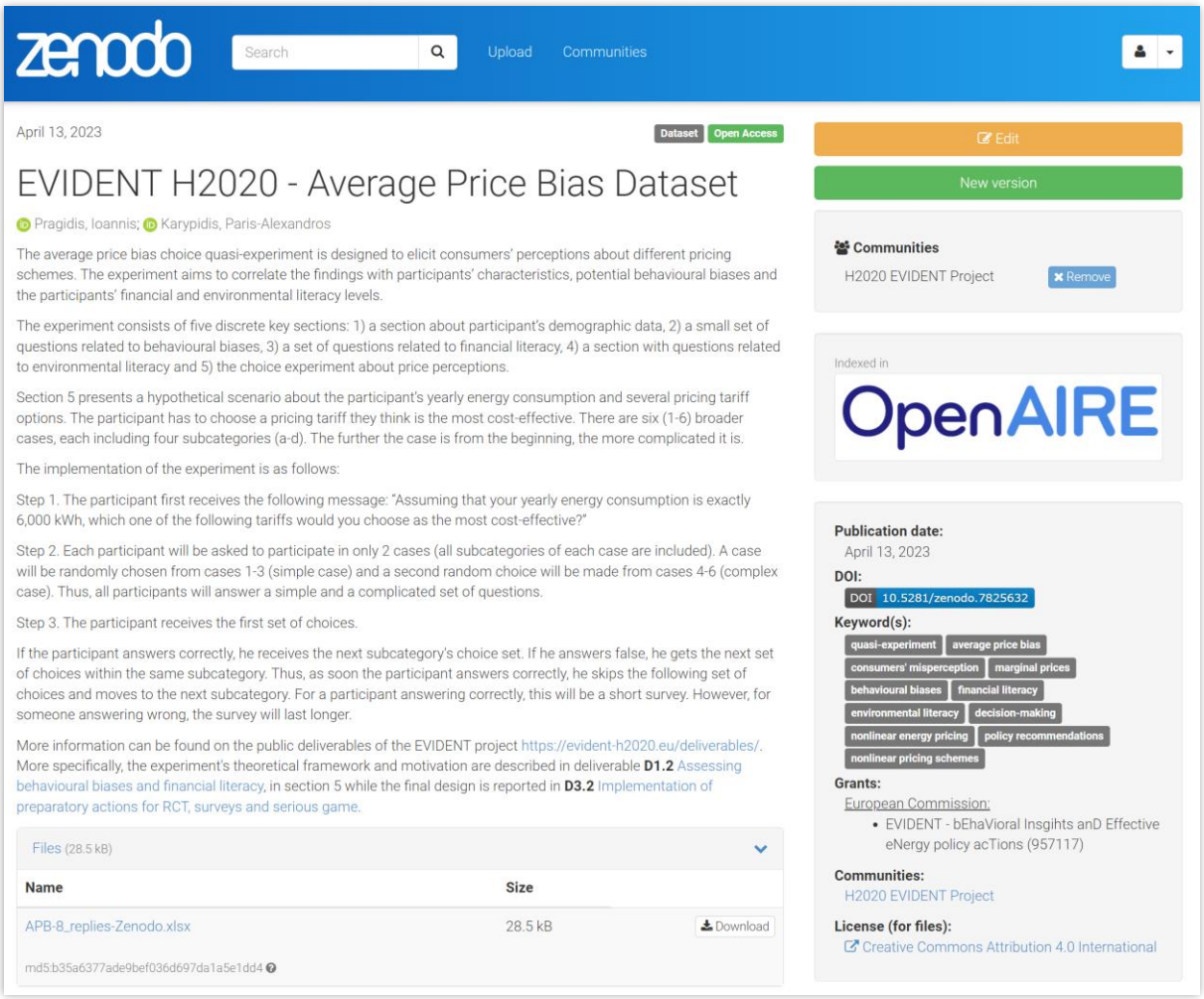


Figure 8: Average Price Bias dataset on Zenodo repository

Figure 9: Discrete Choice Experiment dataset on Zenodo repository

4. Data Access and Sharing

This section will describe the implementation performed on the EVIDENT dataset for access and sharing and a brief summary of the data-sharing requirements. The datasets used in use cases 1 and 2 are not shared or stored on the Zenodo platform. However, the analysis of this data and the knowledge from it will be used as knowledge in the analysis of use cases 4 and 5.

The EVIDENT project will need to identify preferred data-sharing formats to ensure compatibility and interoperability. Open data formats such as CSV or XML are recommended. The data to be shared should meet certain requirements, such as quality control and ethical reviews, including the privacy and confidentiality of participants. A relevant example could be that all data should have metadata. Furthermore, the data to be shared will only come from use cases 4 and 5, and from use case 3 only the meteorological data, including the privacy and confidentiality of participants.

The organizer of the experiments decides the data that will be generated and shared by the EVIDENT platform. The EVIDENT project has procedures for obtaining informed consent from participants and anonymizing and de-identifying data.

4.1 Data Sharing through APIs

An important functionality of the above platforms is the availability of using APIs to extract and process the uploaded research outputs. An API is a set of instructions that allows different software applications to communicate with each other. It provides a way for one program to ask another program for information or data, and then receive that information in a format it can understand.[11]

APIs offer a streamlined and standardized way to access data, allowing researchers to automate data retrieval and analysis, saving time and resources. In this manner, APIs can be used to access large amounts of data from multiple sources, making them a scalable solution for research projects requiring data from multiple sources.

There are already several APIs available for accessing data relevant to EU policy support, such as those provided by Eurostat, the European Data Portal (data on agriculture, education, energy, etc.), European Environment Agency (data on air quality, water quality, greenhouse gas emissions, etc.), European Central Bank (data on exchange rates, interest rates, inflation, etc.), and the European Commission.

The data repositories referenced in previous sections also have APIs. For example, the Zenodo API supports the publishing (deposit) of research outputs, searching published records, and uploading and downloading files (https://developers.zenodo.org). Examples of how to use free and open-source

[11] Vaccari, L., Boyd, M., Posada Sanchez, M., API strategy essentials for Public Sector Innovation: Technical perspective, Publications Office of the European Union, Luxembourg, 2022, doi:10.2760/151339, JRC129951.

tools are provided on the Zenodo website. In this manner, the EVIDENT datasets and research outputs will be accessible and reusable, and compliant with the FAIR principles.[12]

4.1.1 Data space

The European Data Strategy aims to create a single data space by promoting European values and digital rights. The creation of a single data space will ensure that higher volumes of data will be available for use in the economy and society by ensuring that data owners are in control of their data. Leveraging this data space, developers can create applications to be utilised by users. In addition, users will choose how to make their data available, for how long, and under what conditions. Also, they will be able to remove the access permissions. For instance, in the health data space, each patient can share her/his health data with other doctors and hospitals, while ensuring the ownership and control of the data. The nine European Data Spaces are depicted in Figure 10.

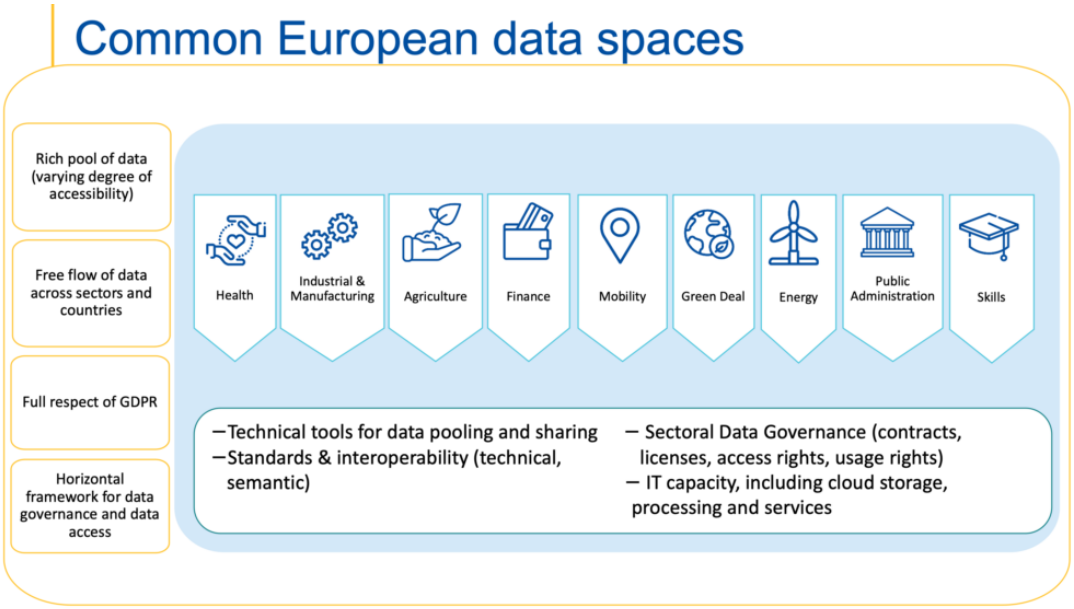


Figure 10: Common European data spaces [21]

4.2 Data Sharing Policies and Data Sharing Agreement

In summary, the policy of free access has been adopted, as well as making the research data available under the Creative Common license. Section 3 presented the dataset to be distributed as open data on the Zenodo platform. Also, each dataset will be assigned a License. Participants in the survey will be informed that their responses will be shared anonymously. In addition, the data generated by the

[12] The use of APIs for data sharing is also foreseen in the EVIDENT grant agreement.

platform will be the property of the organizer of the experiments and the license for the use of the experiments will be held by the organizer of the experiments.

4.3 Data Ethics and Privacy in Horizon 2020 Projects

Research ethics is one of the fundamental principles that underpin Horizon 2020[13]. It is essential to follow ethical guidelines to ensure the protection of human subjects, animals, and the environment. By following ethical guidelines, Horizon 2020 projects can conduct innovative, impactful, socially responsible, and sustainable research.

In particular, Horizon 2020 projects must follow the ethical principles outlined in the EU Charter of Fundamental Rights[14], the Universal Declaration of Human Rights[15], and the Convention on Human Rights and Biomedicine[16]. These principles include respect for human dignity, autonomy, and the right to privacy.

Regarding data protection, researchers must obtain informed consent from human subjects and ensure their confidentiality. For example, Horizon 2020 projects must comply with the EU General Data Protection Regulation (GDPR)[17], which sets out strict rules on how personal data is collected, processed, and stored. Personal data refers to any information that can directly or indirectly identify an individual, such as name, address, email, or IP address.

Informed consent means that participants are fully aware of the research project's purpose, the type of data collected, and how it will be used and protected[18]. Participants must have the right to withdraw their consent at any time and have their data deleted.

Researchers must ensure that personal data is anonymized whenever possible which means that the data is stripped of any identifying information, making it impossible to link it to an individual. Additionally, they must ensure that personal data is stored securely and protected from unauthorized access, loss, or theft. The GDPR requires that personal data be stored for no longer than necessary

[13] European Union. (2016). Ethics Review in Horizon 2020 Projects. Retrieved from https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-review_en.pdf

[14] European Union. (2012). Charter of Fundamental Rights of the European Union. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:12012P/TXT&from=EN

[15] United Nations General Assembly. (1948). Universal Declaration of Human Rights. Retrieved from https://www.un.org/en/universal-declaration-human-rights/

[16] Council of Europe. (1997). Convention for the Protection of Human Rights and Dignity of the Human Being with regard to the Application of Biology and Medicine: Convention on Human Rights and Biomedicine. Retrieved from https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/164

[17] European Union. (2013). General Data Protection Regulation. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN

[18] European Union. (2019). Guidance on the Information to be Provided in the Informed Consent. Retrieved from https://ec.europa.eu/health/sites/default/files/files/eudralex/vol-10/2017_01_26_guideline_informed_consent_en.pdf

and that it is deleted or anonymized once the research project is completed. Data breaches should be reported promptly to the relevant authorities[19].

4.4 Data Ethics and Privacy in the EVIDENT Project

In every step of the EVIDENT project, researchers take into consideration the ethics that are associated with the data collection, storage, and processing, including authorisations or notification requirements. These considerations are in accordance with European and national legislation, the common requirements for Horizon 2020 projects, the research ethics guidelines of each partner organisation, the provisions in EVIDENT's grant agreement, general research ethics practices, etc.

The EVIDENT project has received research ethics clearance from the competent committees of various consortium partners (e.g., Joint Research Centre, Trinity College Dublin) and the Horizon 2020 committee.

An overview of EVIDENT's actions on data protection (e.g., compliance with GDPR), measures to ensure data protection, and the Data Protection Officer's responsibilities are described in D2.3 'Serious Game Implementation Design' (pages 64-67). The requirements and relevant actions are provided in the project's Grant Agreement (e.g., Grant Agreement (Articles 36 and 39) and the activities are monitored by the project's Ethics Manager (EM) and the Ethics and Privacy Committee (EPC). The EM and the EPC warrant that all technical activities, use cases, data management and data processing will be conducted ethically adhering to privacy and regulatory constraints[20].

Moreover, personal data is processed under EU Regulation No 45/2001.[21] In addition to the above and within the context of this deliverable, ethical considerations relevant to using Artificial Intelligence (AI) techniques, especially Machine Learning (ML), are considered. These are described in D1.3 "Specifications of big data analytics".

[19] European Union. (2019). Guidance on Security of Personal Data. Retrieved from https://ec.europa.eu/health/sites/default/files/ehealth/docs/data_protection/guidance_on_security_of_personal_data_en.pdf

[20] Source: EVIDENT grant agreement

[21] Regulation (EC) No 45/2001 of the European Parliament and of the Council of 18 December 2000 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data (OJ L 8.

5. Conclusion

The sharing of data plays a crucial role in facilitating the dissemination of knowledge and the formation of scientific communities, while simultaneously fostering research and innovation. The value of data holds significant importance in enhancing the efficiency and performance of various research processes. By sharing raw data through diverse platforms, their value can be amplified as they are utilized by other researchers. This specific task focuses on the documentation of the data intended for evaluating and assessing policy interventions within the EVIDENT project. Transparency is reinforced by making data comprehensible and openly accessible to the public. Open access to data also facilitates researchers to conduct additional analyses using the same dataset.

The EVIDENT consortium leverages the Zenodo platform and its infrastructure to store the data generated in the five use cases. In addition, any uploaded data adheres to the FAIR principles; also, particular attention has been paid to ensuring the anonymity of the data. Finally, the Creative Commons License is adopted, ensuring that use of the data is free for commercial and non-commercial use, provided the appropriate references are included.

References

[1] European Commission, Directorate-General for Research and Innovation, Kania, K., Bucksch, R., Integration of social sciences and humanities in Horizon 2020 – Participants, budgets and disciplines, Publications Office, 2020. [Online]. Available: https://data.europa.eu/doi/10.2777/141795

[2] J. Eaton, Political Economy: A Marxist Textbook, Revised ed. International Publishers Co, 1966

[3] R. Bärenfänger, "Managing Information Services in the Digital Economy," Ph.D. dissertation, School of Management, Economics, Law, Social Sciences and International Affairs, University of St. Gallen, Bamberg, 2017.

[4]IBM, "What is a Knowledge Graph?" [Online]. Available: https://www.ibm.com/topics/knowledge-graph

[5] "Zenodo - Research. Shared." [Online]. Available: https://about.zenodo.org/principles/

[6] Open Research Europe , "Open Data, Software and Code Guidelines." [Online]. Available: https://open-research-europe.ec.europa.eu/for-authors/data-guidelines#appropriaterepository

[7] Invenio, "Welcome — inveniosoftware.org." [Online]. Available: https://inveniosoftware.org/

[8] Invenio, "Invenio Framework v3." [Online]. Available: https://github.com/inveniosoftware/invenio

[9] Commission of the European Communities, "Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions - Re-use of Public Sector Information: review of Directive 2003/98/EC – [SEC(2009) 597]," 2009. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A52009DC0212

[10] The European Parliament and the Council of the European Union, "Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information," 2019. [Online]. Available: https://eur-lex.europa.eu/eli/dir/2019/1024/oj

[11] Open Source Initiative, "Licenses." [Online]. Available: https://opensource.org/licenses/

[12] D. A. Almeida, G. C. Murphy, G. Wilson, and M. Hoye, "Do Software Developers Understand Open Source Licenses?" IEEE/ACM 25th International Conference on Program Comprehension (ICPC), May 2017, pp. 1–11.

[13] L. Rosen, Open Source Licensing: Software Freedom and Intellectual Property Law. Upper Saddle River, NJ: Prentice Hall PTR., 2004.

[14] G. Haff, "The mysterious history of the MIT License," Apr. 2019. [Online]. Available: https://opensource.com/article/19/4/history-mit-license

[15] A. Murray, "Open Source Licenses: Trends and Predictions," Jan. 2020. [Online]. Available: https://www.mend.io/resources/blog/open-source-licenses-trends-and-predictions-2020/

[16] B. Balter, "Open source license usage on GitHub.com," Mar. 2015. [Online]. Available: https://github.blog/2015-03-09-open-source-license-usage-on-github-com/

[17] European Comission, "European Union Public Licence." [Online]. Available: https://commission.europa.eu/content/european-union-public-licence_en

[18] A. Thevenet, “Slovakia’s Action plan for smart cities and Regions ,” May 2023. [Online]. Available: https://joinup.ec.europa.eu/collection/open-source-observatory-osor/news/slovakias-action-plan-smart-cities-and-regions

[19] G. Shankaranarayanan, R. Y. Wang, and M. Ziad "IP-MAP: Representing the Manufacture of an Information Product," Conference on Information Quality, 2020.

[20] J.D. Gass, A. Misra, M. N. S. Yadav, F. Sana, C. Singh, A. Mankar, B. J. Neal, J. Fisher-Bowman, J. Maisonneuve, M. M. Delaney, K. Kumar, V. P. Singh, N. Sharma, A. Gawande, K. Semrau, and L. R. Hirschhorn, “Implementation and results of an integrated data quality assurance protocol in a randomized controlled trial in Uttar Pradesh, India,” Trials, vol. 18, no. 1. Springer Science and Business Media LLC, Sep. 2017.

[21] Real-time Linked Dataspaces, “Common European Data Spaces.” [Online]. Available: https://dataspaces.info/common-european-data-spaces/

Appendix

CW Dataset

The data are used in use cases 1, 2 and 3. More specifically, in use cases 1 and 2, the data are used in two different cases, to create personalized home energy reports for CW's clients (Figure **12**) and to analyse the effect of the intervention (Figure 14). In addition, the data are used to cover the needs of both analyses in use case 3. The first analysis uses the data to build a residential energy consumption and production forecasting framework (Figure **14**), while the second uses the data to build a methodological framework for estimating heterogeneity in causal effects (Figure 15). Table 9 summarizes the data used in the context of use cases 1, 2 and 3.

Table 9: Four different use cases where CW's data are used

HourValues_Evident_unique (10.000r × 7c)

id	clientid	hour	bought	produced	sold	totalconsumed
0	2.740	2023-02-24 00:00:00	0	2	0	2
0	2.729	2023-02-24 00:00:00	0	2	0	2
0	2.726	2023-02-24 00:00:00	0	0	0	0
0	2.727	2023-02-24 00:00:00	0	2	0	2
0	2.719	2023-02-24 00:00:00	0	1	0	1
0	2.731	2023-02-24 00:00:00	0	0	0	0
0	2.723	2023-02-24 00:00:00	0	5	0	5

Figure 11: Raw hourly data from CW's database

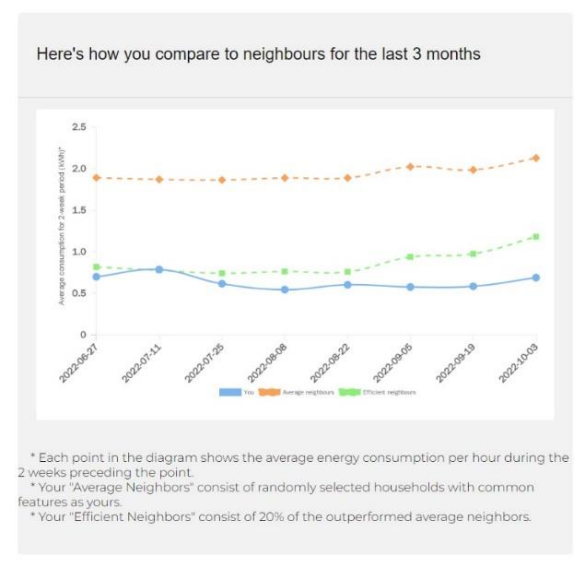


Figure 12: Sample of use case 1 and 2 HERs

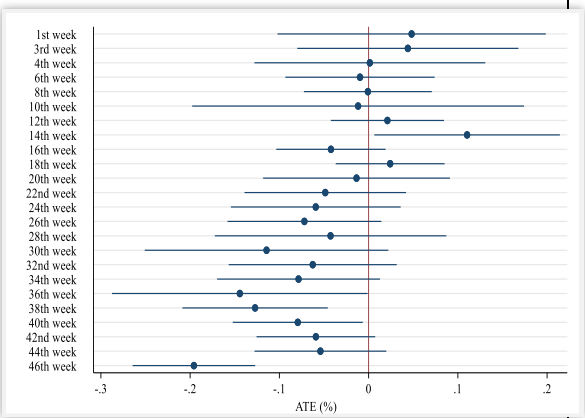


Figure 13: Estimation of intervention's effect on prosumers energy consumption and energy bought

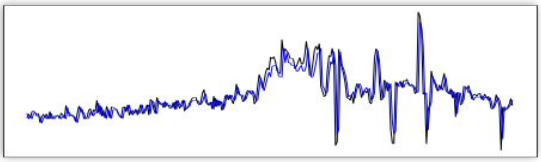


Figure 14: Household energy consumption forecasting

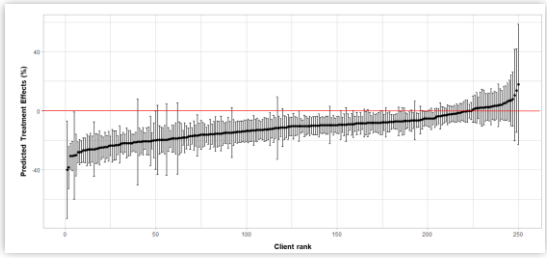


Figure 15: Detection of heterogeneity on randomized control trial

PPC Dataset

The nature of the dataset after pre-processing is mostly categorical with predefined values in each category. This format is not suitable for ML and predictive algorithms since such tools require quantitative data for better exploitation. Therefore, the categorical data are converted (data transformation) into numerical ones using statistical analysis, assumptions and other formulas in the direction of energy efficiency (Table 10).

Table 10:PPC Data Set

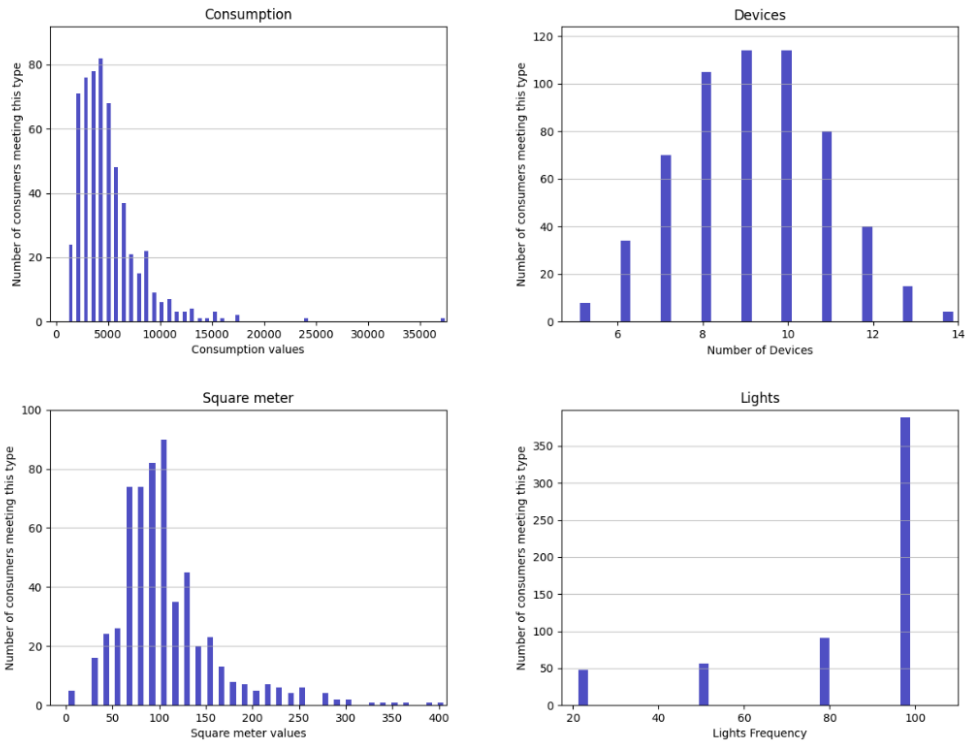


Figure 16: Demographic and energy measurement histograms

Person_who_decides	House_year_built	Windows_panel_type	Floor	Heating_system	Hot_water_system	Lig
16	24	29	40	49	165	60
16	26	29	42	46	53	59
17	25	29	39	165	53	58
15	25	28	42	46	54	59
16	26	29	41	46	54	166
17	24	29	40	69	53	60

Consuption	Sum_of_m2	Device_id	Household_members	Age_group	House_type
3749	208	6.26365e+18	3	8	12
3715	74	6.26365e+19	3	6	12
2752	100	6.26365e+20	3	6	12
4175	101	6.26365e+09	2	6	11
4527	100	6.26365e+15	3	6	12
5471	72	6.26365e+23	4	6	12

Figure 17: Numerical transformation of the categorical dataset along with consumptions

Environmental Dataset for Use Case 3

Apache Spark is used to manage the data, which also assists in the creation of the relevant attributes. Figure 18 shows the source code of the agent created for the meteorological data collection, while Figure 19 shows the data stored in the database by executing a query to have an overview of the data schema and also to see some indicative values. Figure 20 shows the data aggregated by each city, while Figure 21 depicts the consolidated data by Day Section. The Day Section column is the result of dividing the day into 3 sections depending on the time of the measurement. This approach is used to achieve dimensionality reduction and aggregation of measurements. Also, the text description from the weather became vectors, and then a new column, "weather type", was created. In addition, sunrise and sunset were used to get the duration in seconds of the sun in the city. Moreover, Figure 22 shows the data with the problematic timestamps. Initially, the UNIX time format was used to form the timestamp, but then changed to another format, while the timestamps in version 4 were removed. Figure 23 presents the Swedish cities and the averages from the meteorological data for the whole period of the measurements. It should be mentioned that up to version 3 in the dataset, the file is not an open file format, while in version 4 it was converted to PDF; also, a CSV format was added. Figure 24 presents the average temperatures for all months, while Figure 25 shows a monthly comparison of temperatures between Stockholm and Gothenburg. It should be mentioned that the comparison of temperatures in the day is made with the Day Section column, meaning that for each month, the averages of the Day Section are obtained. Furthermore, Figure 26 presents an instance of the number of measurements per Day Section Alunda city on the 1st , 2nd , and 3rd of April. Finally, Figure 27, shows the monthly average temperature values per day section with the average day length in hours.

Figure 18: Software agent to gather meteorological data from the Swedish cities

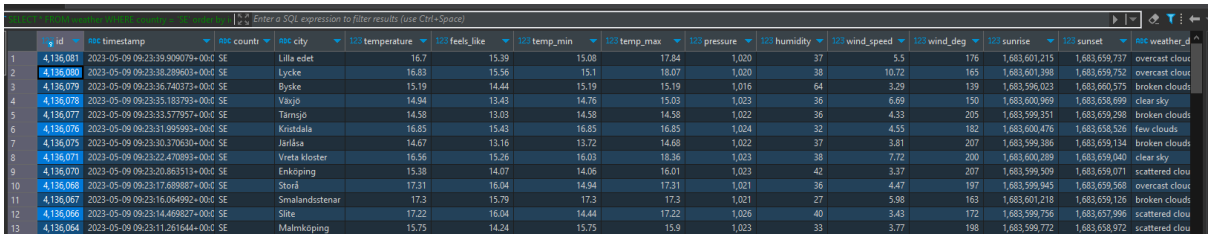
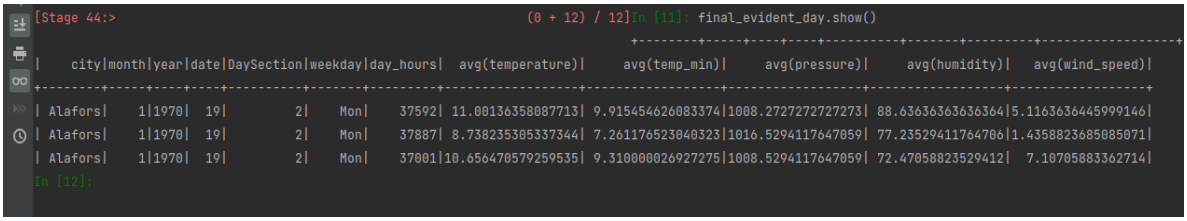
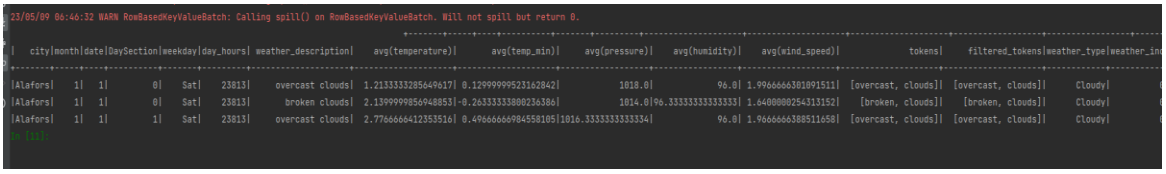


Figure 19: PostgreSQL raw data



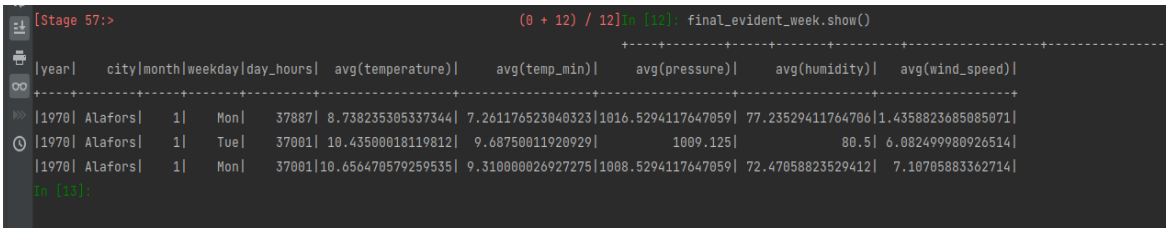
```
[Stage 44:>                                                 (0 + 12) / 12]In [11]: final_evident_day.show()
+--------+-----+----+----+----------+-------+---------+------------------+------------------+------------------+------------------+------------------+
|    city|month|year|date|DaySection|weekday|day_hours|  avg(temperature)|     avg(temp_min)|     avg(pressure)|     avg(humidity)|   avg(wind_speed)|
+--------+-----+----+----+----------+-------+---------+------------------+------------------+------------------+------------------+------------------+
| Alafors|    1|1970|  19|         2|    Mon|    37592| 11.00136358087713| 9.915454626083374|1008.2727272727273| 88.63636363636364|5.1163636445999146|
| Alafors|    1|1970|  19|         2|    Mon|    37887| 8.738235305337344| 7.261176523040323|1016.5294117647059| 77.23529411764706|1.4358823685085071|
| Alafors|    1|1970|  19|         2|    Mon|    37001|10.656470579259535| 9.310000026927275|1008.5294117647059| 72.47058823529412|  7.10705883362714|
+--------+-----+----+----+----------+-------+---------+------------------+------------------+------------------+------------------+------------------+
In [12]:
```

Figure 20: Average day-section



```
23/05/09 06:46:32 WARN RowBasedKeyValueBatch: Calling spill() on RowBasedKeyValueBatch. Will not spill but return 0.
+------+-----+----+---------+-------+--------+--------------------+------------------+------------------+------------------+------------------+------------------+--------------------+--------------------+------------+
|  city|month|date|DaySection|weekday|day_hours| weather_description|  avg(temperature)|     avg(temp_min)|     avg(pressure)|     avg(humidity)|   avg(wind_speed)|              tokens|    filtered_tokens|weather_type|
+------+-----+----+---------+-------+--------+--------------------+------------------+------------------+------------------+------------------+------------------+--------------------+--------------------+------------+
|Alafors|    1|   1|        0|    Sat|   23813|    overcast clouds| 1.2133333285649617| 0.1299999952316284|            1018.0|              96.0| 1.9966666301091511|  [overcast, clouds]|  [overcast, clouds]|      Cloudy|
|Alafors|    1|   1|        0|    Sat|   23813|      broken clouds| 2.1399999856948853|-0.2633333380023638|            1014.0|96.33333333333333| 1.6400000254313152|    [broken, clouds]|    [broken, clouds]|      Cloudy|
|Alafors|    1|   1|        1|    Sat|   23813|    overcast clouds| 2.7766666412353516| 0.4966666698455811|1016.3333333333334|              96.0| 1.9666666388511658|  [overcast, clouds]|  [overcast, clouds]|      Cloudy|
In [13]:
```

Figure 21:Day-section and weather description



```
[Stage 57:>                                                 (0 + 12) / 12]In [12]: final_evident_week.show()
+----+--------+-----+-------+---------+------------------+------------------+------------------+------------------+------------------+
|year|    city|month|weekday|day_hours|  avg(temperature)|     avg(temp_min)|     avg(pressure)|     avg(humidity)|   avg(wind_speed)|
+----+--------+-----+-------+---------+------------------+------------------+------------------+------------------+------------------+
|1970| Alafors|    1|    Mon|    37887| 8.738235305337344| 7.261176523040323|1016.5294117647059| 77.23529411764706|1.4358823685085071|
|1970| Alafors|    1|    Tue|    37001| 10.43500018119812|  9.68750011920929|          1009.125|             80.5| 6.082499980926514|
|1970| Alafors|    1|    Mon|    37001|10.656470579259535| 9.310000026927275|1008.5294117647059| 72.47058823529412|  7.10705883362714|
+----+--------+-----+-------+---------+------------------+------------------+------------------+------------------+------------------+
In [13]:
```

Figure 22: Timestamp error

	DaySection 0						DaySection 1					DaySection 2				
Cities	Average temperature	Average humidity	Average temp_min	Average pressure	Average wind_speed	Average temperature	Average humidity	Average temp_min	Average pressure	Average wind_speed	Average temperature	Average humidity	Average temp_min	Average pressure	Average wind_speed	Total Average temperature
Åby	4.618115848	79.26519	3.67019068	1013.335	3.571375476	7.421454077	69.20197482	6.36659745	1012.806	4.426006624	3.754874671	82.32891	2.846381055	1013.163	3.173930196	4.770903121
Åhus	6.419768466	86.19836	5.81146438	1014.76	3.714075297	8.946802355	76.21839519	8.30038159	1014.299	4.352513038	5.580946892	88.98179	4.884727697	1014.757	3.285902561	6.525608276
Åkersberg	4.480762957	80.28213	3.28196712	1012.85	4.11763977	6.787226617	73.47336134	5.67525759	1012.618	4.233774507	3.614887349	83.00132	2.481997611	1012.914	3.799173694	4.528901093
Alafors	6.185595359	82.51993	4.66194811	1013.203	2.63816345	8.418989003	74.26326281	6.87167774	1012.733	2.873408231	5.458454248	84.25623	4.000834512	1013.06	2.53368367	6.285689364
Älberga	4.365160809	73.92114	3.52976463	1013.408	3.316952294	6.879176521	69.19370864	6.19147025	1012.87	3.525678323	3.426085849	82.07381	2.574393895	1013.279	3.020608514	4.423854192
Åled	6.005558556	84.26137	5.44548673	1013.916	4.165924745	8.141543115	75.46508287	7.58436854	1013.484	4.820805775	5.850420186	85.12088	5.07671851	1013.728	3.849507698	6.399825727
Alfta	3.10374194	79.76562	3.0624074	1012.11	3.037258097	5.010568466	74.74243574	4.96440278	1011.775	3.267045215	1.513395709	86.2435	1.49474731	1012.472	2.712168928	2.658425285
Älgarås	4.30267296	83.80374	3.36828575	1013.432	3.624405509	6.684153433	76.18170996	5.82173663	1012.806	3.774970014	3.86419443	86.8218	2.766667029	1013.225	3.321991527	4.595102863
Alingsås	5.615642494	76.97317	4.32892847	1013.554	3.455206559	8.034283185	70.16073285	6.66740535	1012.99	3.707701104	4.99822306	84.22407	3.773880913	1013.598	3.121787087	5.82174574
Almeboda	5.044194597	75.74751	4.24332592	1014.659	3.541286803	7.562076966	68.03997146	6.65384721	1014.033	3.768287632	4.149141568	84.837	3.300139864	1014.727	3.163901577	5.122177286
Almhult	5.587610136	81.77278	5.45182028	1014.675	3.602279696	7.812498057	73.50120788	7.72662226	1013.955	3.852395124	4.254225126	86.32018	4.139768936	1014.542	3.228995922	5.364342588
Almunge	3.538574658	74.96035	3.03920784	1012.698	3.368659011	6.236581892	69.30941846	5.6220942	1012.377	3.495265304	2.823207212	82.47645	2.184261162	1012.779	3.027576856	3.751403804
Älta	4.628295689	80.73652	3.15873782	1010.669	2.981686445	6.773285078	72.33733417	5.48001192	1010.459	3.585734282	3.638877523	83.60003	2.261643566	1010.553	2.528363085	4.572080576
Alunda	4.422479581	75.45072	4.2469727	1012.301	4.332190511	5.717333278	71.15426471	5.53592986	1012.133	4.327347673	3.584348743	80.74073	3.437095243	1012.447	3.99805046	4.250269735
Alvängen	6.218519716	82.72082	4.75723065	1013.137	2.599290522	8.523464508	74.72870215	7.06760565	1012.822	2.804914588	5.466839274	84.41756	4.157270028	1013.009	2.475229378	6.322432263
Alvesta	5.311999917	85.58296	4.60330579	1013.821	3.515871605	8.051540236	72.13426937	7.15812366	1013.309	4.038102076	4.549791381	88.4889	3.682630283	1013.787	3.080284804	5.510765662
Alvsjö	4.648232604	80.97006	3.29461827	1010.272	2.896722605	6.842384531	72.71715892	5.671857	1009.818	3.487989663	3.653706125	83.5382	2.354915692	1010.227	2.448882255	4.589788641
Åmål	4.706975149	76.88839	4.62593029	1012.986	4.876168683	6.362515227	72.20537223	6.29436908	1012.454	4.88220298	4.45206889	80.4913	4.383670197	1012.819	4.840494361	4.938177288
Åmotfors	3.812844812	77.04721	3.64879632	1012.856	2.782543838	6.322565084	70.78716235	6.20453096	1012.17	3.028263736	2.965036933	84.25331	2.783264656	1012.419	2.556673169	3.920596515
Aneby	4.165803855	77.26899	4.16580386	1013.879	3.556407304	6.774239259	70.75146977	6.77423926	1013.207	3.688397547	3.78739527	85.27106	3.78739527	1013.971	3.225698512	4.543856367
Ängelholm	6.554814993	77.55795	5.48786906	1014.476	4.562583085	8.782501432	69.62294721	7.59965627	1014.154	5.399699448	6.168272211	78.76558	5.050100782	1014.284	4.325704561	6.844012898
Ängelsber	3.976915289	75.57781	3.59322099	1012.773	3.173277113	6.412623543	69.57419581	6.06063245	1012.368	3.329013812	2.909739383	82.91861	2.548166544	1012.764	2.876763982	3.927740622
Angered	5.916366202	83.51314	4.61264586	1012.659	2.980533772	8.090391014	75.40985896	6.69957945	1012.524	3.08711878	5.359147904	85.33633	4.042150516	1012.45	2.939453932	6.094484431
Ankarsrun	5.783483086	74.07674	5.78348309	1013.99	3.459950077	7.453456321	68.14928805	7.45345632	1013.443	3.588334248	4.2058426	82.67308	4.2058426	1013.884	3.092917964	5.289715691
Arboga	5.237433783	75.4279	3.43052134	1013.135	3.578546007	7.658113379	70.1158672	6.23793968	1012.672	3.67409087	4.249910421	82.08585	2.866248916	1013.164	3.421615643	5.236057823

Figure 23: Average per day-section

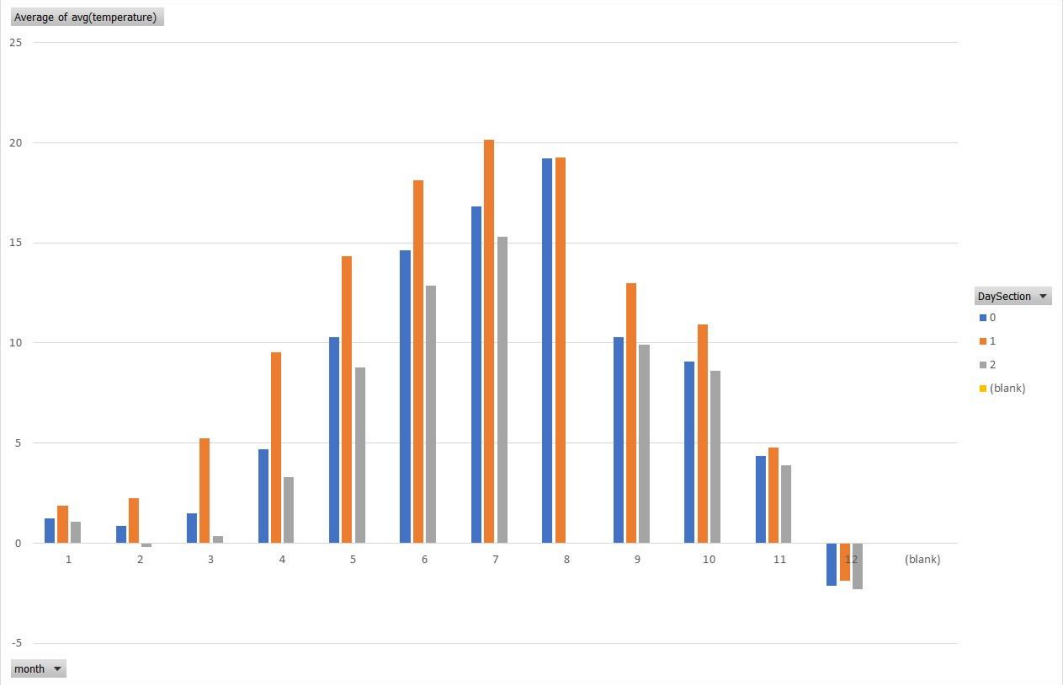


Figure 24: Average of temperature

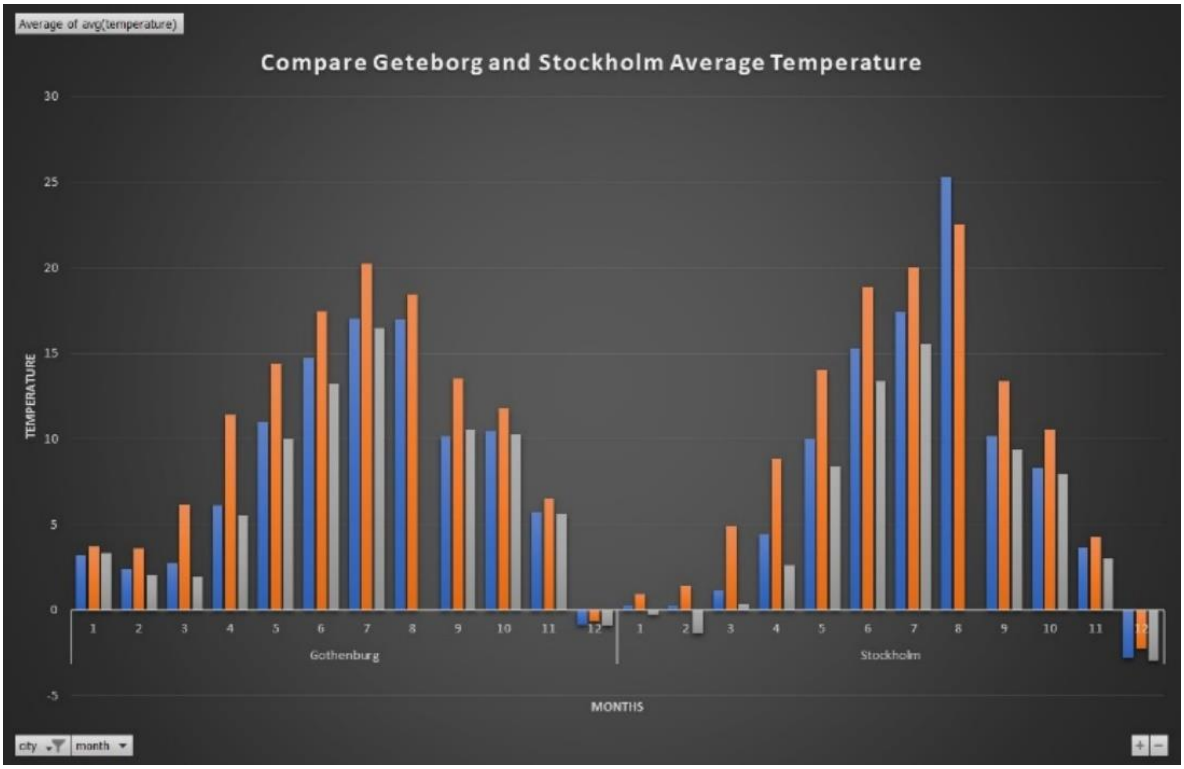


Figure 25: Monthly Compare Cities temperature in Day section

city	Alunda		
month	4		
Count of temperature	Years		
	2022	2023	Grand Total
1	3	5	8
0	1	1	2
1	1	1	2
2	1	3	4
2	2	5	7
0		1	1
1	1	1	2
2	1	3	4
3	4	5	9
0	1	1	2
1	1	1	2
2	2	3	5

Figure 26: example of number of measurements in Aluna city of 1, 2, and 3 April

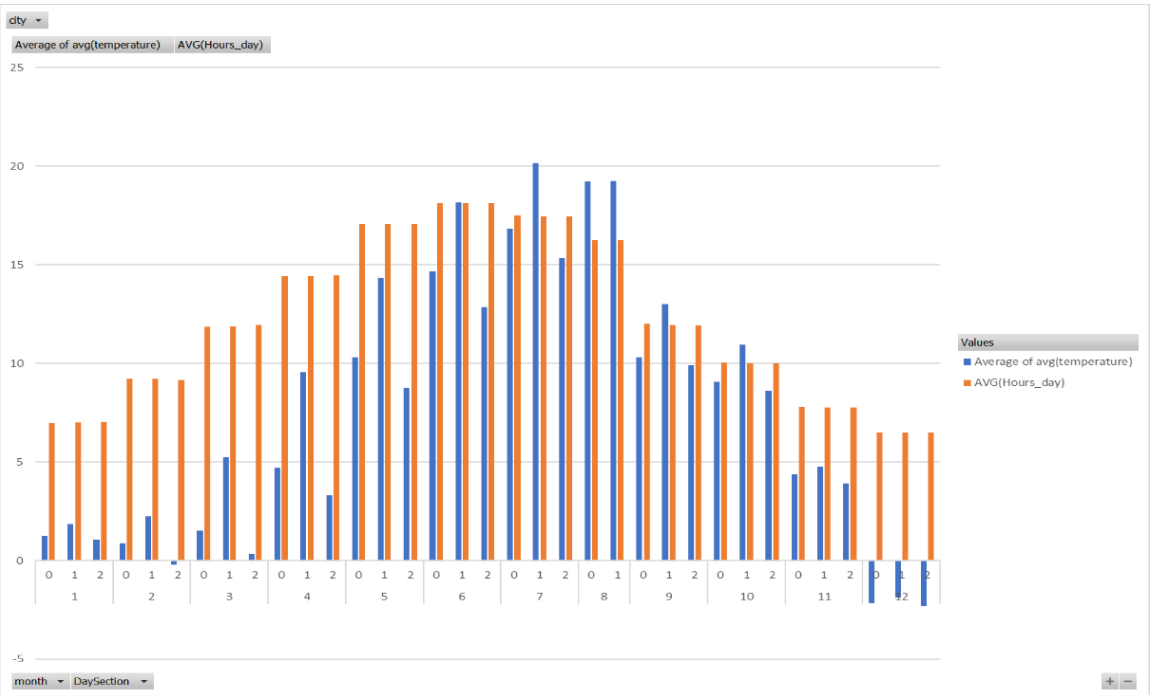


Figure 27: Monthly average Hours od day and temperature.