Enabling a Data-informed Public Sector: An Introductory Course to BDTI Essentials

Session 2: Data Cleaning and Transformation

Welcome!
The webinar will begin shortly.

DG CNECT
Directorate-General for Communications
Networks, Content and Technology

DG DIGIT
Directorate-General
for Digital Services

Some housekeeping

- The Chat and Q&A widgets are located to the right of the presenter screen
- Please post session-related questions in Q&A
- Microphones and video are off by default, attendees cannot turn them on
- The session will be recorded and available shortly after the end of the session on our website

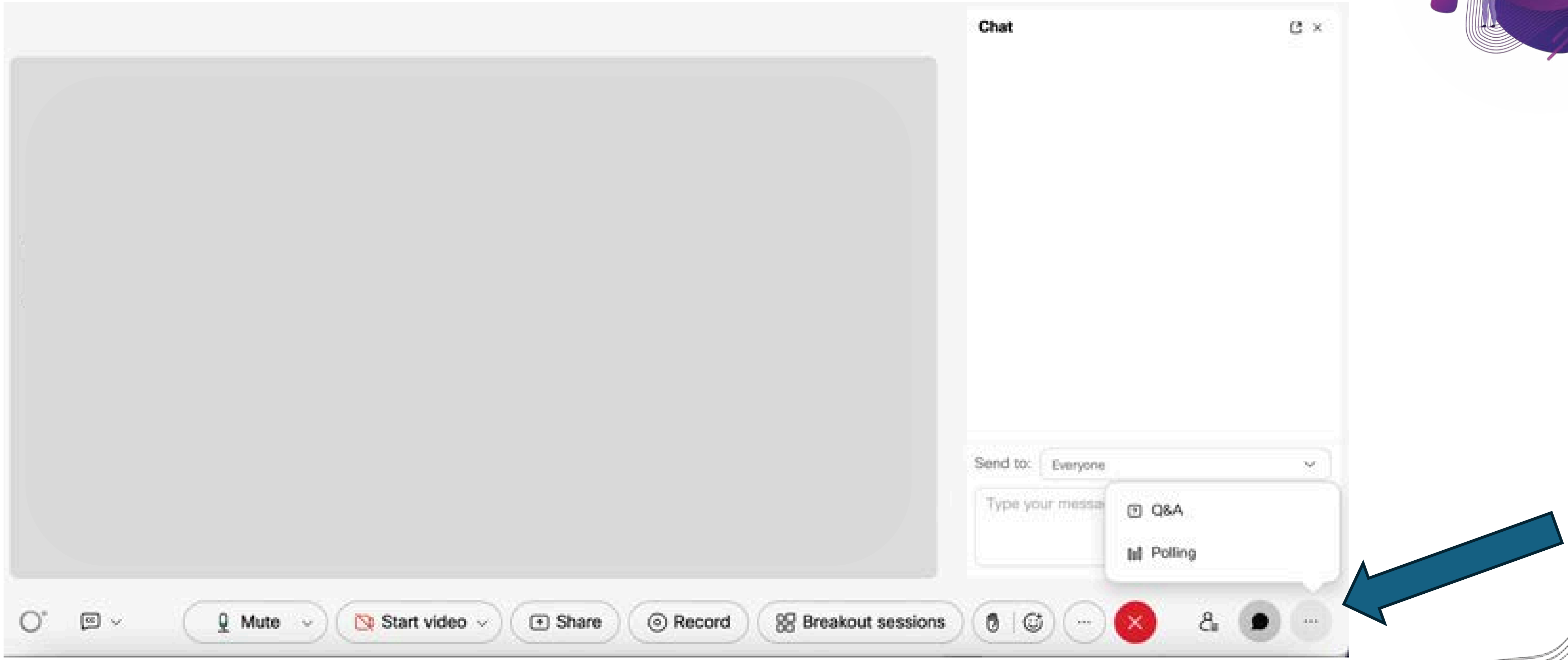
Do

- Participate in the chat during the session
- Ask questions during the Q&A portion at the end
- Give feedback after the session
- Join the dedicated discussion board

Don’t

- Self-promote
- Disrespect anyone

Locate Chat and Q&A - Webex



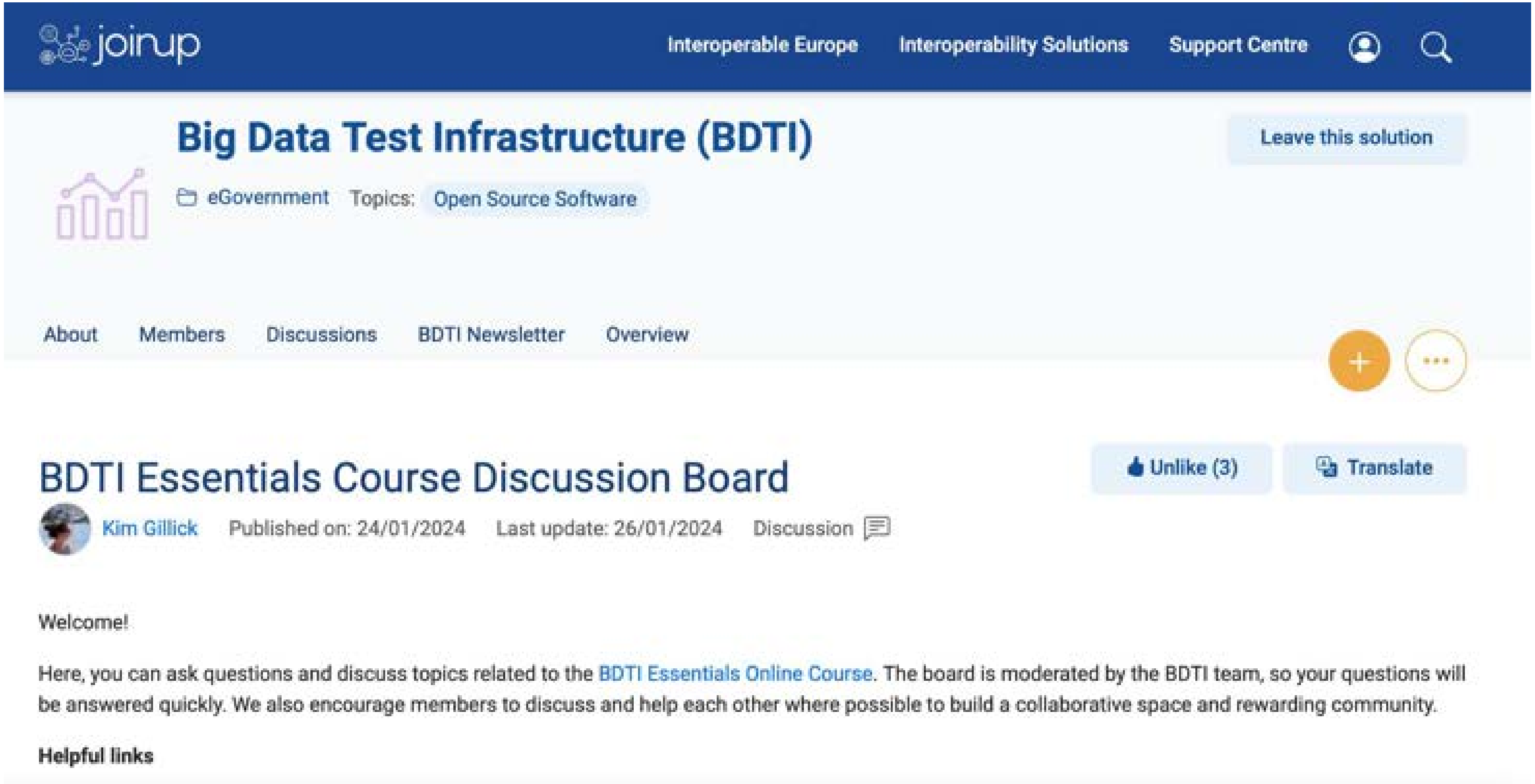
Enabling a Data-informed Public Sector: An Introductory Course to BDTI Essentials

Session 2: Data Cleaning and Transformation

Agenda

- Welcome [11:00-11:05]
- Session 1 Recap [11.05-11.15]
- Data cleaning and transformation [11:15-12:00]
- Q&A [12:00-12:10]
- Wrap up [12:10 – 12:15]

Course discussion board



Big Data Test Infrastructure (BDTI)

BDTI Essentials Course Discussion Board

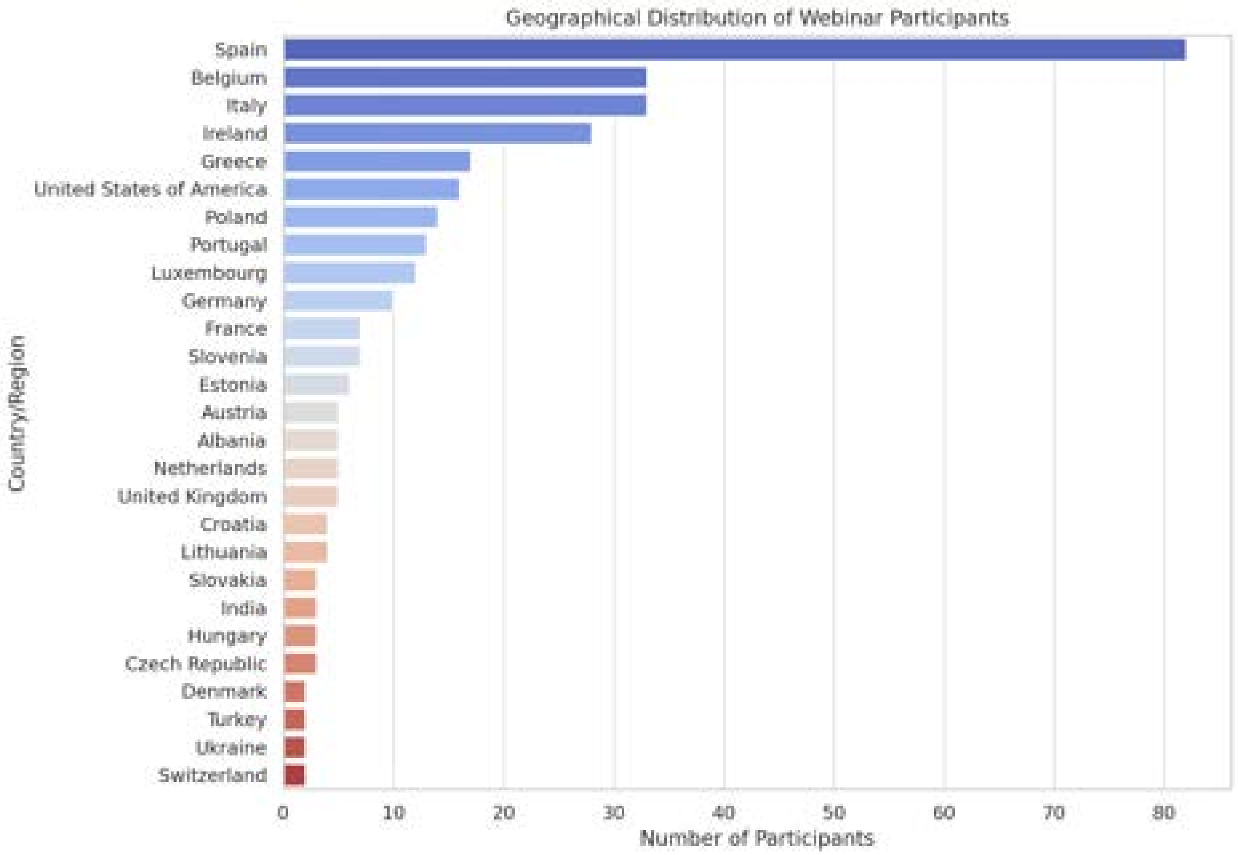
Kim Gillick   Published on: 24/01/2024   Last update: 26/01/2024   Discussion

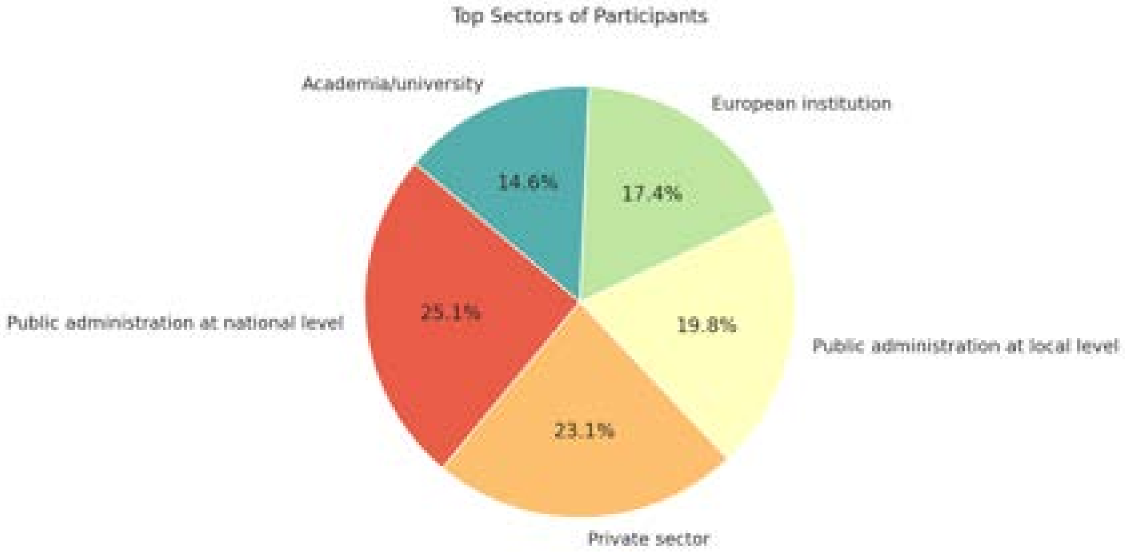
Welcome!

Here, you can ask questions and discuss topics related to the BDTI Essentials Online Course. The board is moderated by the BDTI team, so your questions will be answered quickly. We also encourage members to discuss and help each other where possible to build a collaborative space and rewarding community.

Helpful links

A bit about you





Session 1 –
BDTI intro,
Data Access and Exploration Recap and Exercise review

European
Commission

What is the Big Data Test Infrastructure (BDTI) ?

Not **only** for big data, for **public sector in general (open data)**

You have the key ingredients (datasets),
we equip you with the best **open-source tool**
to create amazing recipes for **public good.**

Six months free of charge service
for EU public administrations *

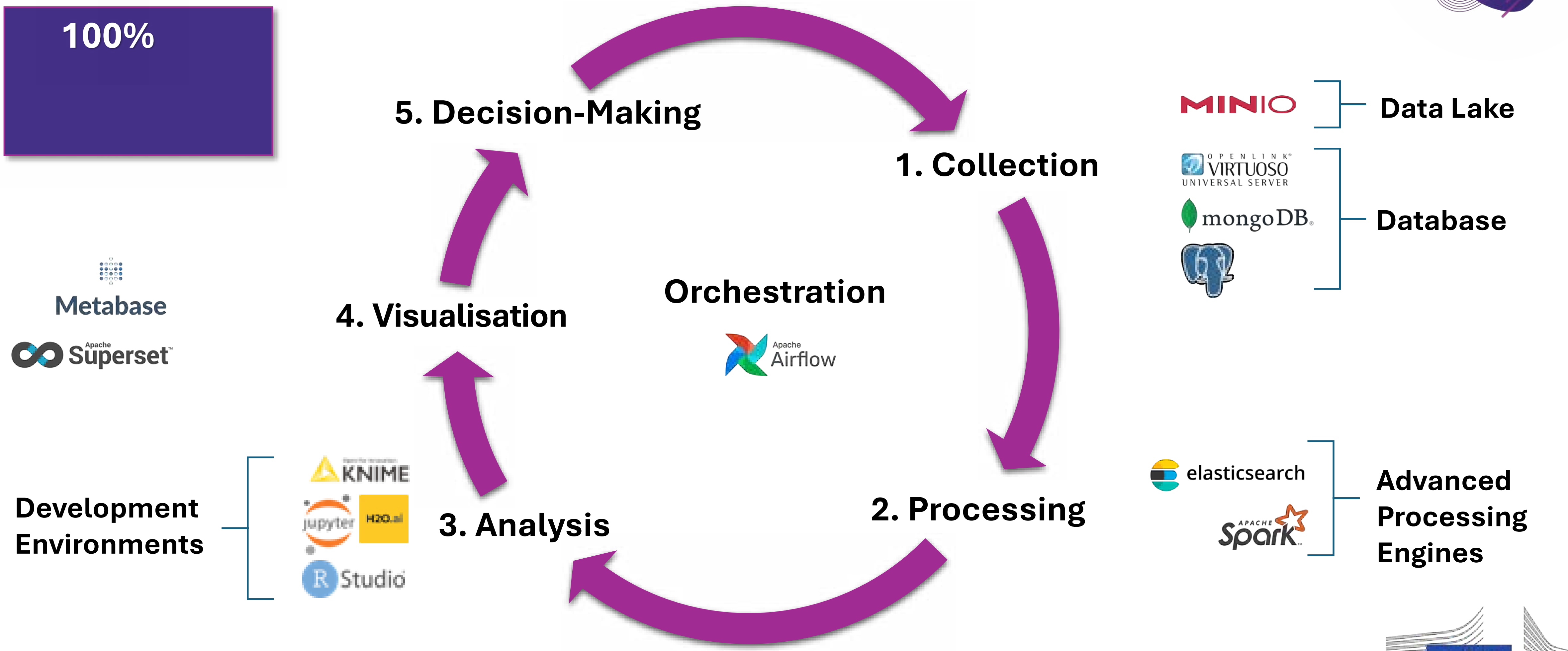
Ready-to-use
data analytics stack and
support

Cloud platform based on
open-source tools

* The cost of the pilot project must fit within the funding boundaries of the BDTI pilot budget

Open-source tools to support your data journey

100%



Who is the Big Data Test Infrastructure (BDTI) for?

European Public Administrations
All European Public Administrations at **local, regional and national level** can independently apply for a BDTI pilot project

Ecosystem with academia and private sector
Academia, spin-off, startups can apply for pilot projects once there is a **clear collaboration** with a Public Administration which will be the main point of contact for the project (**Master/PhD, GovTech startups)**

Are you working for a public administration in need of infrastructure for data analytics?

https://big-data-test-infrastructure.ec.europa.eu/

European Commission

Use case:
"Identify Innovative Green Energy Projects from EU Universities"

Zoi and her team, work in the **Educational Department of a highly polluted EU region.**

Their **purpose** is to identify potential ***"green energy partnerships"*** between EU universities to address pollution issues.

They found some interesting **open data:**

- **List of innovation *green energy* projects** across EU countries
- **CO2 emissions** of these countries

Zoi reached out the BDTI team to learn how to **connect the dots**, use the free BDTI playground to **create *insights*** and produce a ***report***, which can be **easily updated** when new data is available.

The Data Analysis Process is like cooking a dish

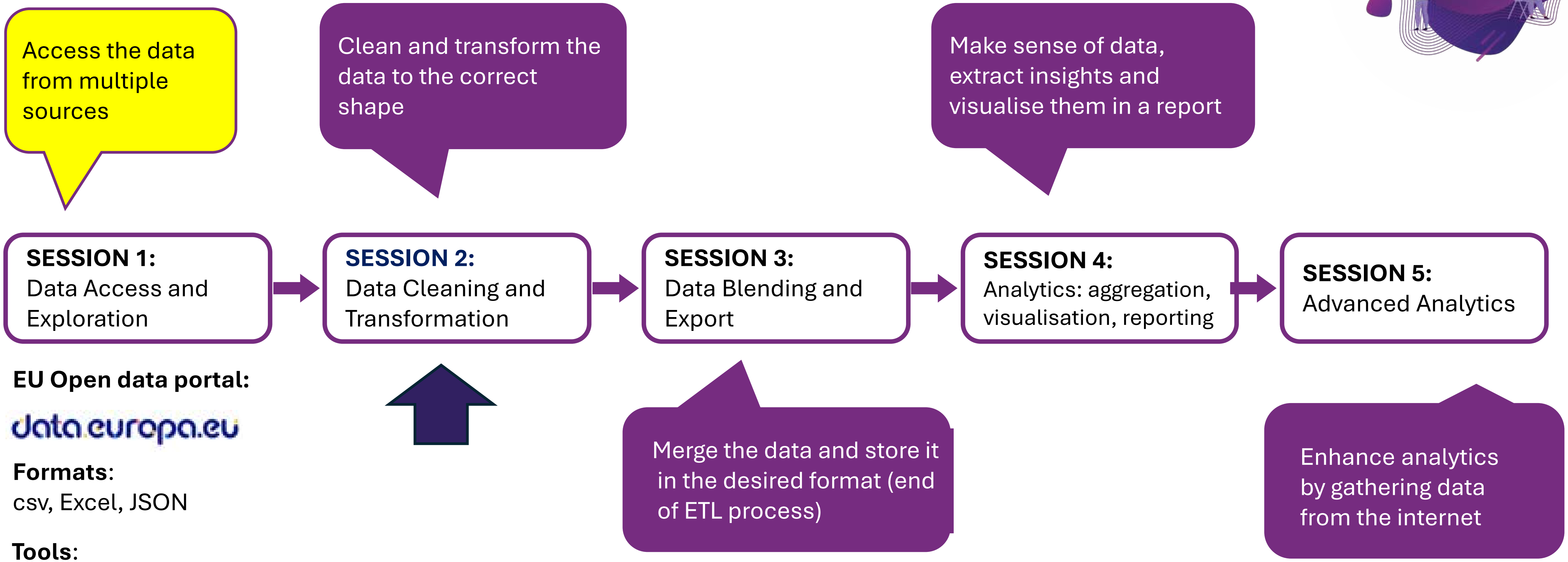
Business challenge/question/problem statement

Ingredients + tools

1. **Find** the data you need → *gather the right ingredients, good quality*
2. **Get, clean** and **prepare** your data → *slice and dice*
3. **Analyse** your data → *mix ingredients together and try different combinations*
4. **Present** the results and **create** knowledge → *serve and consume*

Data→ Information → Knowledge

Recap: Zoi Data Journey



Access the data from multiple sources

Clean and transform the data to the correct shape

Make sense of data, extract insights and visualise them in a report

SESSION 1:
Data Access and Exploration

SESSION 2:
Data Cleaning and Transformation

SESSION 3:
Data Blending and Export

SESSION 4:
Analytics: aggregation, visualisation, reporting

SESSION 5:
Advanced Analytics

Merge the data and store it in the desired format (end of ETL process)

Enhance analytics by gathering data from the internet

EU Open data portal:

data.europa.eu

Formats:
csv, Excel, JSON

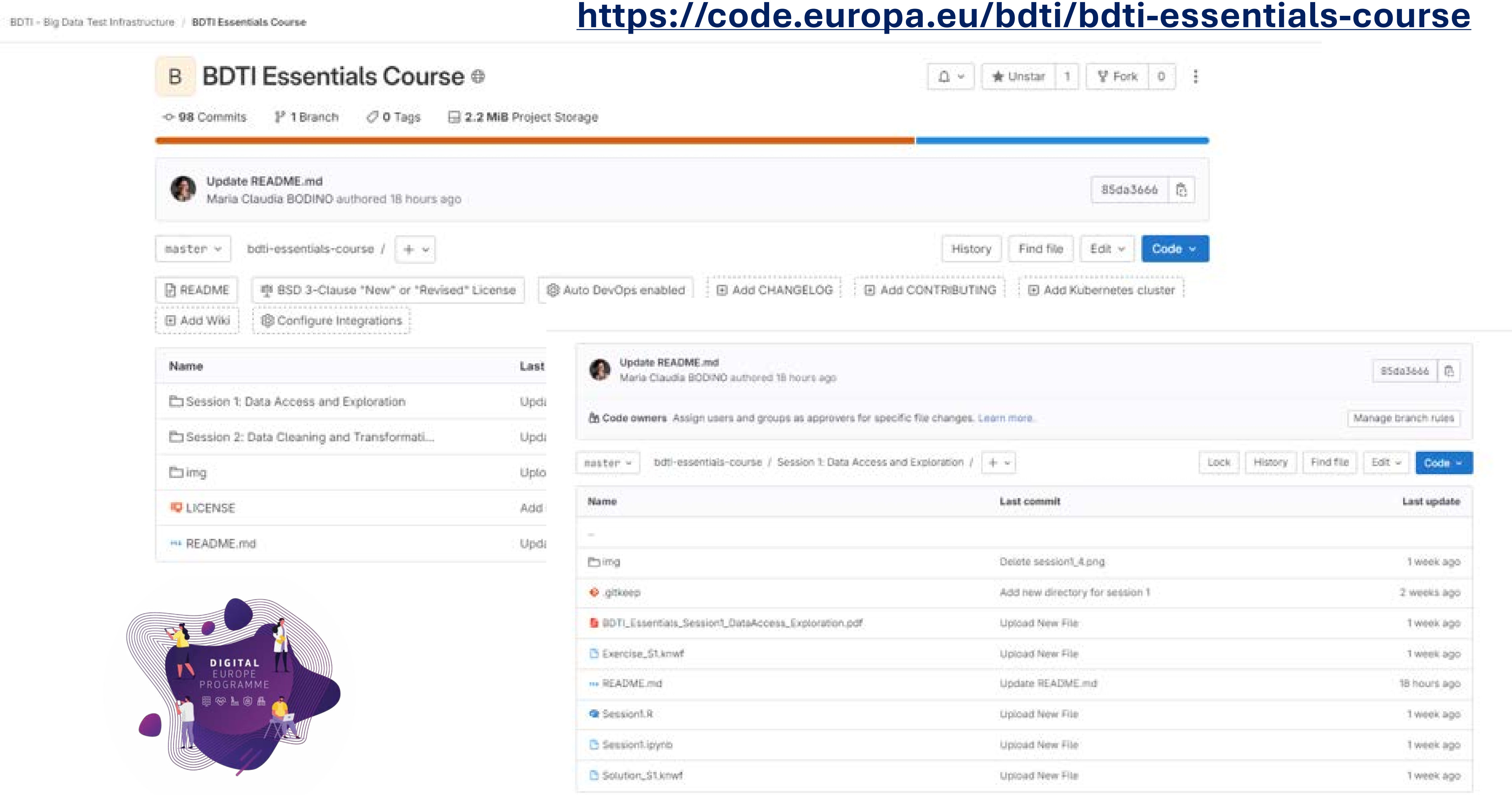
Tools:

Open for Innovation KNIME

jupyter

RStudio

https://code.europa.eu/bdti/bdti-essentials-course



BDTI - Big Data Test Infrastructure / BDTI Essentials Course

B BDTI Essentials Course

Unstar 1 | Fork 0

98 Commits | 1 Branch | 0 Tags | 2.2 MiB Project Storage

Update README.md
Maria Claudia BODINO authored 18 hours ago
85da3666

master | bdti-essentials-course / | + | History | Find file | Edit | Code

README | BSD 3-Clause "New" or "Revised" License | Auto DevOps enabled | Add CHANGELOG | Add CONTRIBUTING | Add Kubernetes cluster | Add Wiki | Configure Integrations

Name	Last
Session 1: Data Access and Exploration	Upda
Session 2: Data Cleaning and Transformati...	Upda
img	Uplo
LICENSE	Add
README.md	Upda

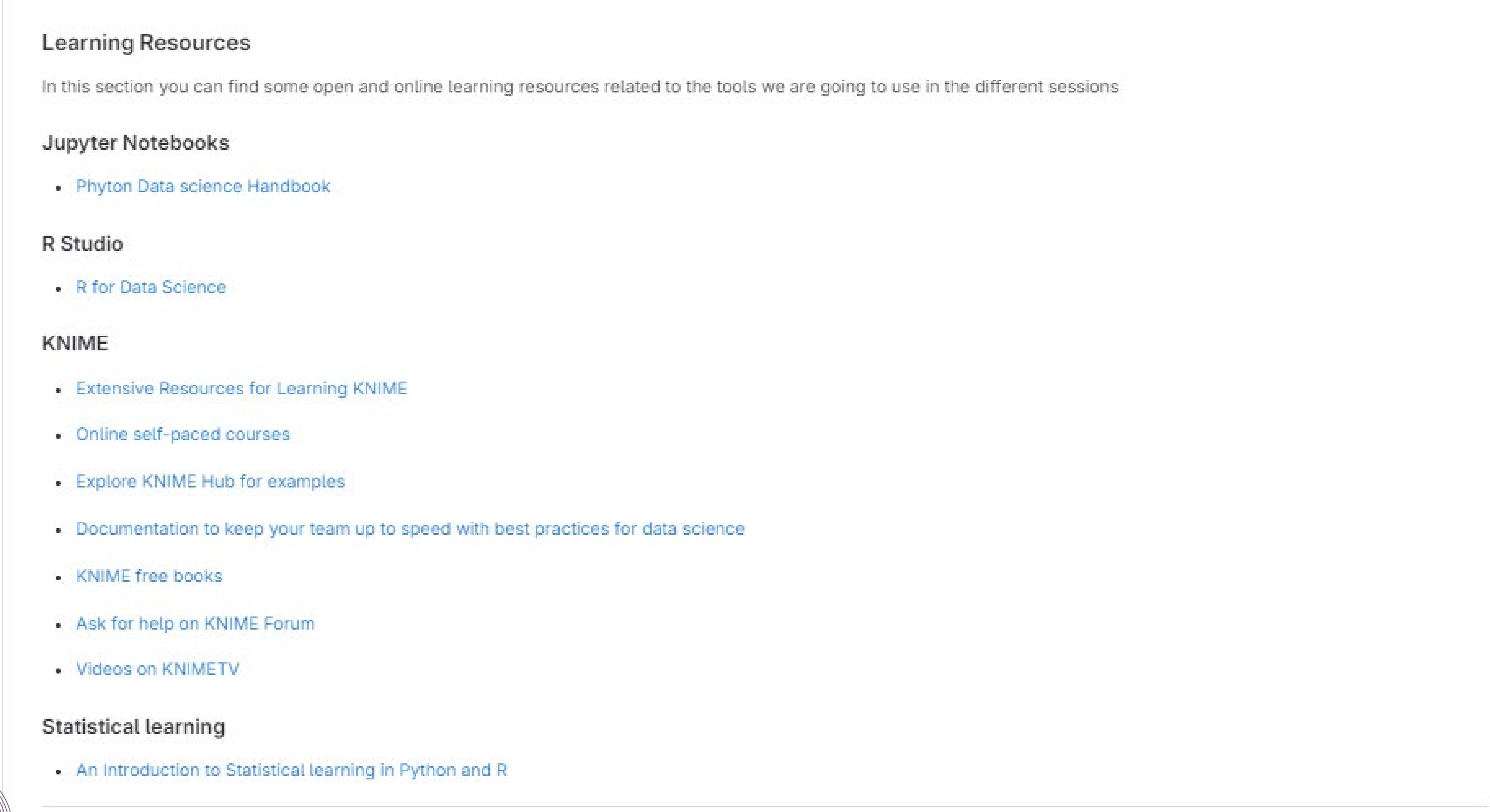
Update README.md
Maria Claudia BODINO authored 18 hours ago
85da3666

Code owners Assign users and groups as approvers for specific file changes. Learn more. | Manage branch rules

master | bdti-essentials-course / Session 1: Data Access and Exploration / | + | Lock | History | Find file | Edit | Code

Name	Last commit	Last update
..		
img	Delete session1_4.png	1 week ago
.gitkeep	Add new directory for session 1	2 weeks ago
BDTI_Essentials_Session1_DataAccess_Exploration.pdf	Upload New File	1 week ago
Exercise_S1.knwf	Upload New File	1 week ago
README.md	Update README.md	18 hours ago
Session1.R	Upload New File	1 week ago
Session1.ipynb	Upload New File	1 week ago
Solution_S1.knwf	Upload New File	1 week ago

DIGITAL EUROPE PROGRAMME



Learning Resources

In this section you can find some open and online learning resources related to the tools we are going to use in the different sessions

Jupyter Notebooks

- Phyton Data science Handbook

R Studio

- R for Data Science

KNIME

- Extensive Resources for Learning KNIME
- Online self-paced courses
- Explore KNIME Hub for examples
- Documentation to keep your team up to speed with best practices for data science
- KNIME free books
- Ask for help on KNIME Forum
- Videos on KNIMETV

Statistical learning

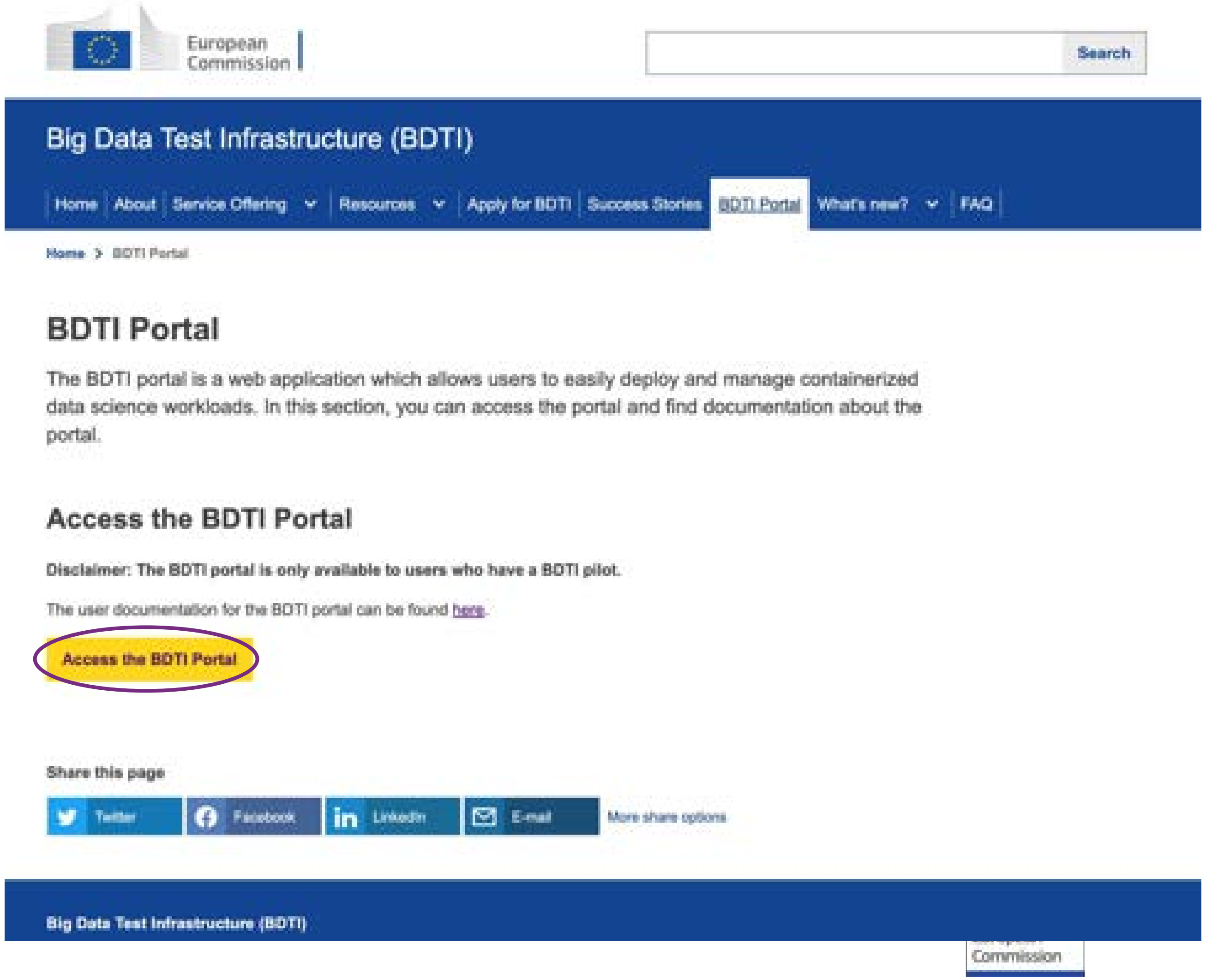
- An Introduction to Statistical learning in Python and R

Exercise Review

BDTI access refresh

- To have access, public administrations need to ***apply on the official website***
- https://big-data-test-infrastructure.ec.europa.eu/apply-bdti_en

- Once you have access follow these steps:
  - Step 1 > Access the portal via the website
  - Step 2 > Navigate to the "Service Catalog" and launch the tools
  - Step 3 > Go to "My Services" and open the tool you created



Big Data Test Infrastructure (BDTI)

BDTI Portal

The BDTI portal is a web application which allows users to easily deploy and manage containerized data science workloads. In this section, you can access the portal and find documentation about the portal.

Access the BDTI Portal

Disclaimer: The BDTI portal is only available to users who have a BDTI pilot.

The user documentation for the BDTI portal can be found here.

Share this page

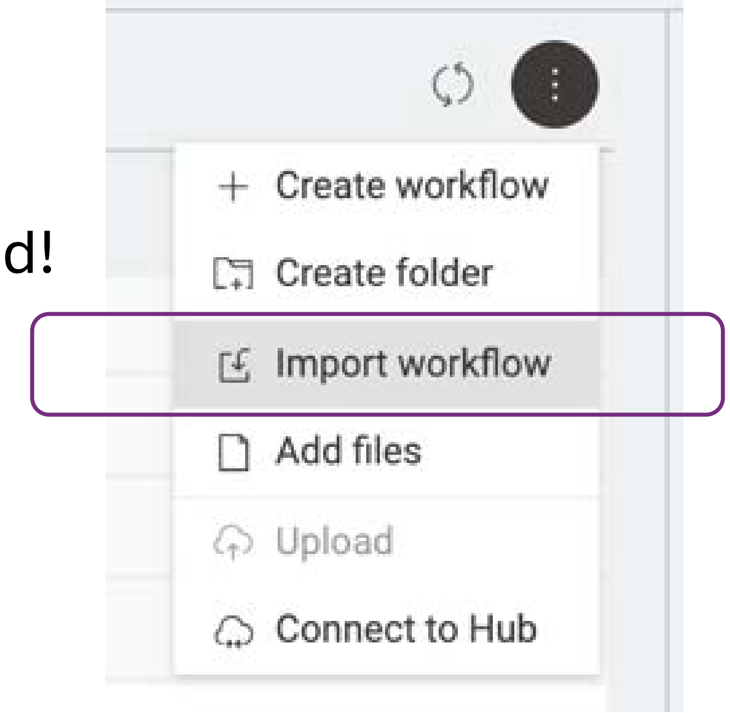
Big Data Test Infrastructure (BDTI)

Exercise review

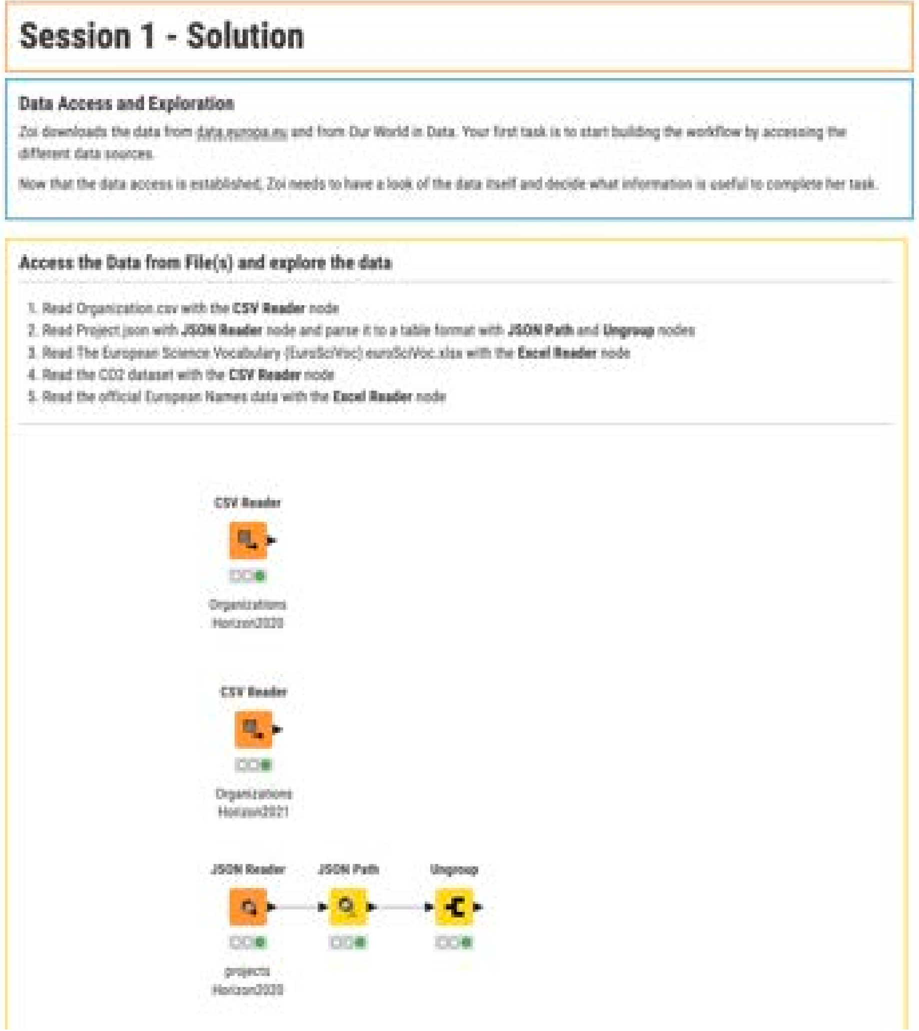
How to import an existing KNIME workflow?

- Once you are in your working space click on "import workflow"
  - .knwf = 'workflow'
  - .knar = 'archive'

- Remember to reconfigure when needed!



Exercise review



Session 1 - Solution

Data Access and Exploration

Zoi downloads the data from data.europa.eu and from Our World in Data. Your first task is to start building the workflow by accessing the different data sources.

Now that the data access is established, Zoi needs to have a look of the data itself and decide what information is useful to complete her task.

Access the Data from File(s) and explore the data

1. Read Organization.csv with the **CSV Reader** node
2. Read Project.json with **JSON Reader** node and parse it to a table format with **JSON Path** and **Ungroup** nodes
3. Read The European Science Vocabulary (EuroSciVoc) euroSciVoc.xlsx with the **Excel Reader** node
4. Read the CO2 dataset with the **CSV Reader** node
5. Read the official European Names data with the **Excel Reader** node

Session 2:
Data Cleaning and Transformation

Session 2: Data Cleaning and Transformation

Session	Topic	Duration
Webinar 1	Data Access and Exploration	45 min
Webinar 2	**Data Cleaning and Transformation**	**75 min**
Webinar 3	Data Blending and Storage	75 min
Webinar 4	Analytics	75 min
Webinar 5	Advanced Module: Gathering Data from the Web and Geo Visualisation	75 min

- Slides, hands-on exercises, and solutions will be provided on the webinar
- Gitlab repository https://code.europa.eu/bdti/bdti-essentials-course

Recap of objectives

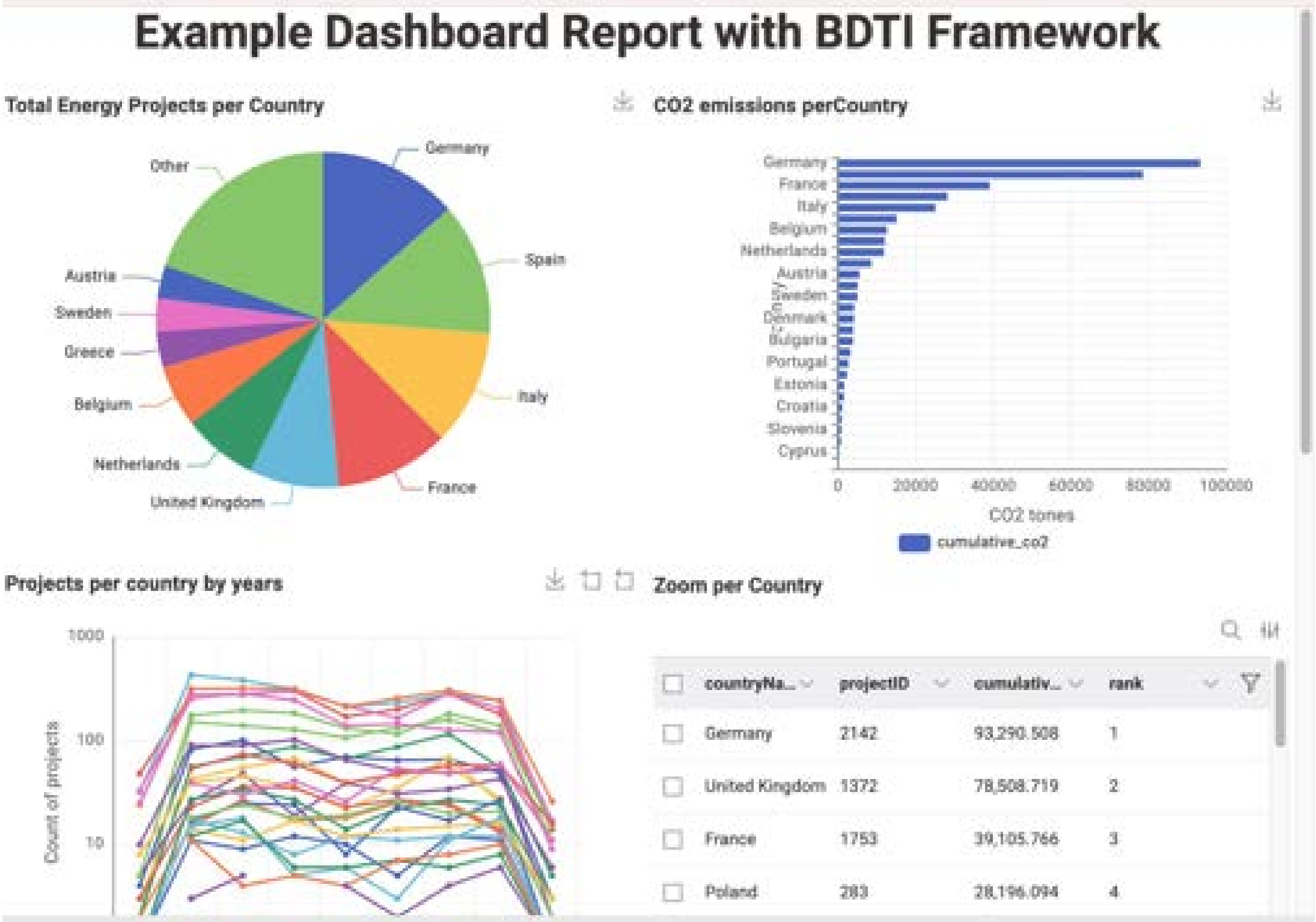
Objective	Level
Visualise total **Energy** projects	EU Countries
	Organisations in EU Countries
Identify and **Visualise** CO2 emissions	EU Countries
Visualise energy projects **per year**	EU Countries
Explore the Relationship between CO2 emissions and Energy Project Funding	EU Countries
Map Visualisation of funded projects	EU Universities

Outputs according to Zoi goals

DIGITAL EUROPE PROGRAMME

The report/dashboard needs to be informative!

- Insightful plots
- Tables with a rank of countries according to CO2 emissions
- Bar chart with CO2 emissions
- Line plot with both Horizon 2020 and 2021
- Locate on a map the Universities that got funding for Energy projects from Horizon



Example Dashboard Report with BDTI Framework

Total Energy Projects per Country

CO2 emissions perCountry

Projects per country by years

Zoom per Country

□	countryNa...	projectID	cumulativ...	rank
□	Germany	2142	93,290.508	1
□	United Kingdom	1372	78,508.719	2
□	France	1753	39,105.766	3
□	Poland	283	28,196.094	4

Commission

Session 2: Learning outcomes

At the end of this session, you should be able to:

- Sign in into the BDTI framework and initialise KNIME.

- Prepare the data for analysis by **cleaning** the dataset.

- Prepare the data for subsequent analysis by **transforming** the dataset to the correct format.

- Address data quality issues.

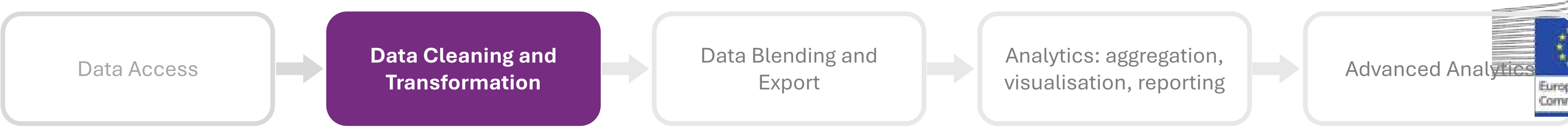
Session 2: Data Cleaning and Transformation

- The **core** of the data pipeline
- Why clean and transform?
  - Data are **rarely** clean
  - **Remove** not useful data
  - Remove **repeated** data
  - Handle **missing values**
- **Extract** more information
  - Define rules and mathematical operations
  - Transform at cell or row **level**
- Transform the data to the desired **shape**
  - Rename and resort table columns
  - Split and merge columns
  - Convert data types

Use Case

With an overview of the data, Zoi realises that it is not in the best shape to do an analysis, some columns are in the wrong format and she is also missing data.

In this session, she will learn techniques to clean and transform the data.



Data Access → **Data Cleaning and Transformation** → Data Blending and Export → Analytics: aggregation, visualisation, reporting → Advanced Analytics

Data Cleaning

Data Cleaning

Why?

- "Dirty data"
  - □ Unnecessary rows or columns

Item	Cost	Store Type	City	VAT	Delivery code
Table	50	Onsite	NY	14%	XY
Couch	100	Online	NY	14%	XX
Couch	100	Online	NY	14%	XX
Cancelled	?	?	?	?	?
TV	70	Online	NY	14%	XY
Chair	30	Onsite	NY	14%	XY
Table	?	Online	NY	14%	XX

Data Cleaning

Why?

- "Dirty data"
  - □ Unnecessary rows or columns
  - □ Missing or null values

Item	Cost	Store Type	City	VAT	Delivery code
Table	50	Onsite	NY	14%	XY
Couch	100	Online	NY	14%	XX
Couch	100	Online	NY	14%	XX
Cancelled	?	?	?	?	?
TV	70	Online	NY	14%	XY
Chair	30	Onsite	NY	14%	XY
Table	?	Online	NY	14%	XX

Data Cleaning

Why?

- "Dirty data"
  - □ Unnecessary rows or columns
  - □ Missing or null values
  - □ Duplicates

Item	Cost	Store Type	City	VAT	Delivery code
Table	50	Onsite	NY	14%	XY
Couch	100	Online	N-Y	14%	XX
Couch	100	Online	N-Y	14%	XX
Cancelled	?	?	?	?	?
TV	70	Online	NY	14%	XY
Chair	30	Onsite	NY	14%	XY
Table	?	Online	NY	14%	XX

Data Cleaning

Why?

- Keep only what is essential

Item	Cost	Store Type	City
Table	50	Onsite	NY
Couch	100	Online	NY
Couch	100	Online	NY
TV	70	Online	NY
Chair	30	Onsite	NY
Table	50	Online	NY

Data Cleaning

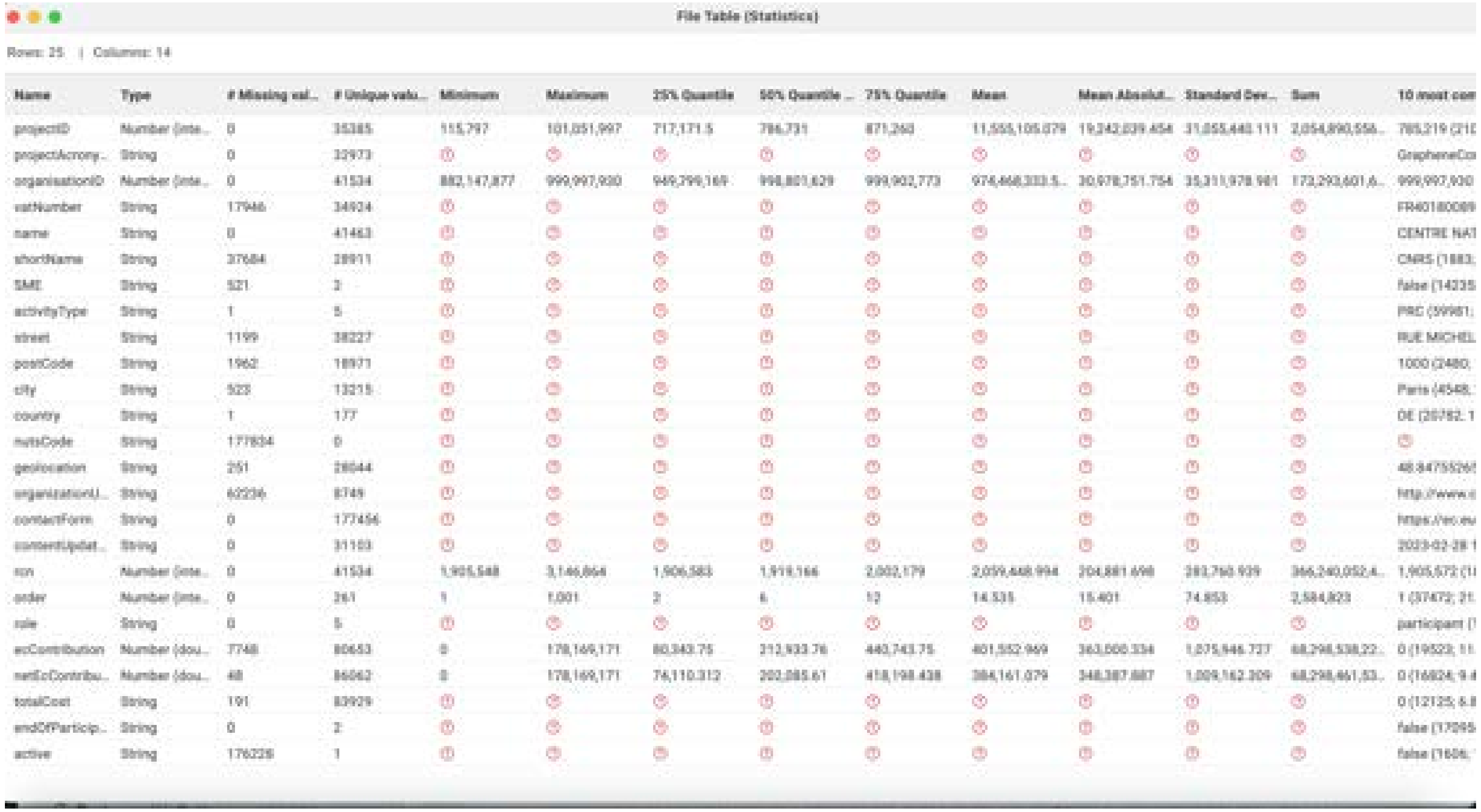
Step 1: Explore data

Step 2: Identify which columns / rows you need and filter or sort

Step 3: Identify if there are missing values or duplicates you need to deal with

Step 1 - Explore Data

- Look at raw data (R-studio, Jupyter Notebook, KNIME)
- Review statics view (KNIME)



File Table (Statistics)

Rows: 25 | Columns: 14

Name	Type	# Missing val…	# Unique valu…	Minimum	Maximum	25% Quantile	50% Quantile …	75% Quantile	Mean	Mean Absolut…	Standard Dev…	Sum	10 most com
projectID	Number (inte…	0	35385	115,797	101,051,997	717,171.5	786,731	871,260	11,555,105.079	19,242,039.454	31,055,440.111	2,054,890,556…	785,219 (210
projectAcrony…	String	0	32973	?	?	?	?	?	?	?	?	?	GrapheneCo
organisationID	Number (inte…	0	41534	882,147,877	999,997,930	949,799,169	998,801,629	999,902,773	974,468,333.5…	30,978,751.754	35,311,978.981	173,293,601,6…	999,997,930
vatNumber	String	17946	34924	?	?	?	?	?	?	?	?	?	FR40180089
name	String	0	41463	?	?	?	?	?	?	?	?	?	CENTRE NAT
shortName	String	37684	28911	?	?	?	?	?	?	?	?	?	CNRS (1883;
SME	String	521	2	?	?	?	?	?	?	?	?	?	false (14235
activityType	String	1	5	?	?	?	?	?	?	?	?	?	PRC (39981;
street	String	1199	38227	?	?	?	?	?	?	?	?	?	RUE MICHEL
postCode	String	1962	18971	?	?	?	?	?	?	?	?	?	1000 (2480;
city	String	523	13215	?	?	?	?	?	?	?	?	?	Paris (4548;
country	String	1	177	?	?	?	?	?	?	?	?	?	DE (20782; 1
nutsCode	String	177834	0	?	?	?	?	?	?	?	?	?	?
geolocation	String	251	28044	?	?	?	?	?	?	?	?	?	48.84755265
organizationU…	String	62236	8749	?	?	?	?	?	?	?	?	?	http://www.c
contactForm	String	0	177456	?	?	?	?	?	?	?	?	?	https://ec.eu
contentUpdat…	String	0	31103	?	?	?	?	?	?	?	?	?	2023-02-28 1
rcn	Number (inte…	0	41534	1,905,548	3,146,864	1,906,583	1,919,166	2,002,179	2,039,448.994	204,881.698	283,760.939	366,240,052,4…	1,905,572 (1
order	Number (inte…	0	261	1	1,001	2	6	12	14.535	15.401	74.853	2,584,823	1 (37472; 21
role	String	0	5	?	?	?	?	?	?	?	?	?	participant (1
ecContribution	Number (dou…	7748	80653	0	178,169,171	80,343.75	212,933.76	440,743.75	401,552.969	363,000.334	1,075,946.727	68,298,538,22…	0 (19523; 11
netEcContribu…	Number (dou…	48	86062	0	178,169,171	74,110.312	202,085.61	418,198.438	384,161.079	348,387.887	1,009,162.209	68,298,461,53…	0 (16824; 9.4
totalCost	String	191	83929	?	?	?	?	?	?	?	?	?	0 (12125; 6.8
endOfParticip…	String	0	2	?	?	?	?	?	?	?	?	?	false (17095
active	String	176228	1	?	?	?	?	?	?	?	?	?	false (1606;

Data Cleaning

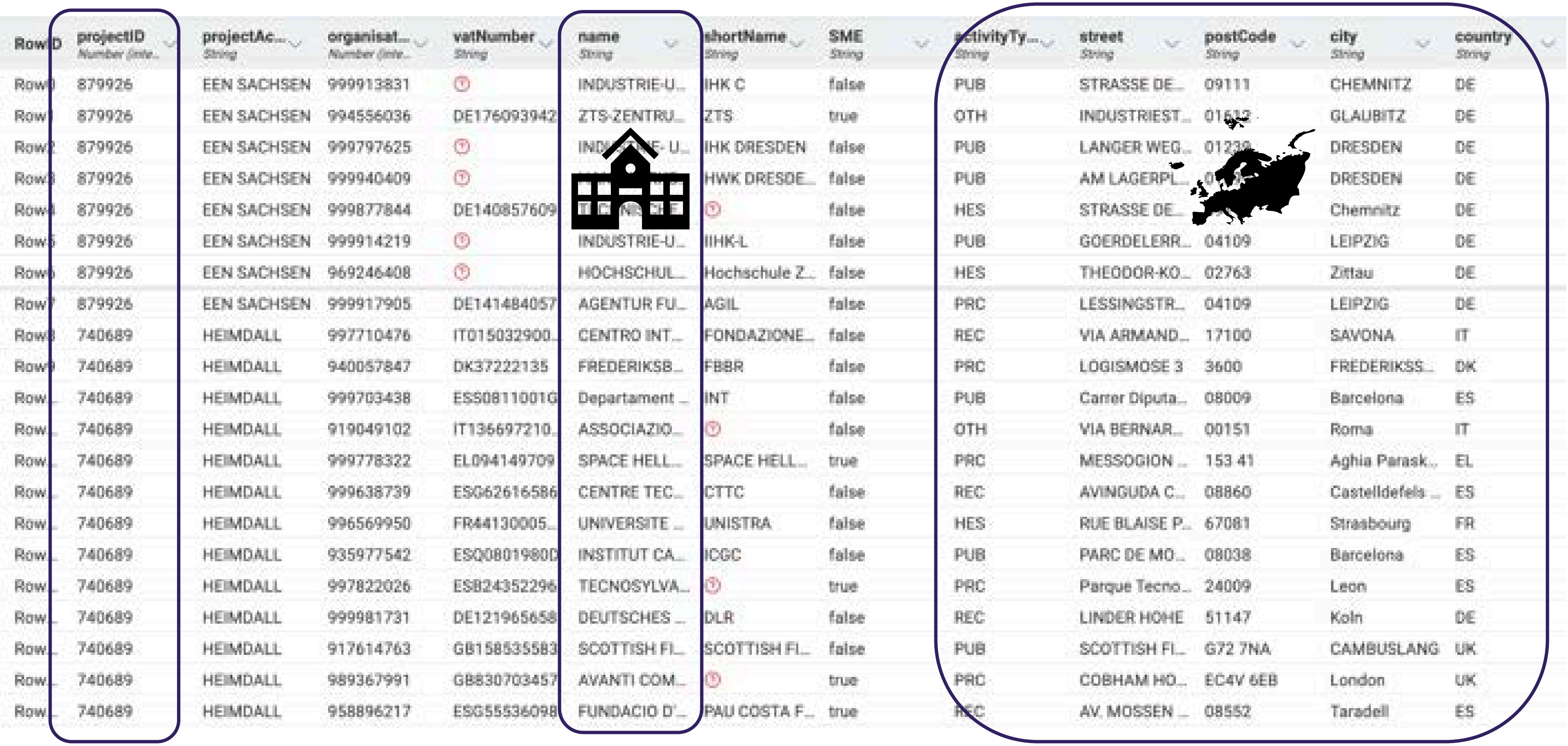
Step 1: Explore data

Step 2: Identify which columns / rows you need, and filter or sort

Step 3: Identify if there are missing values or duplicates you need to deal with

Step 2 - **Identify** columns

- Horizon Data "Organisations"- *Visualise Energy projects per EU Country and for* ***Organisations*** *per* ***Country***

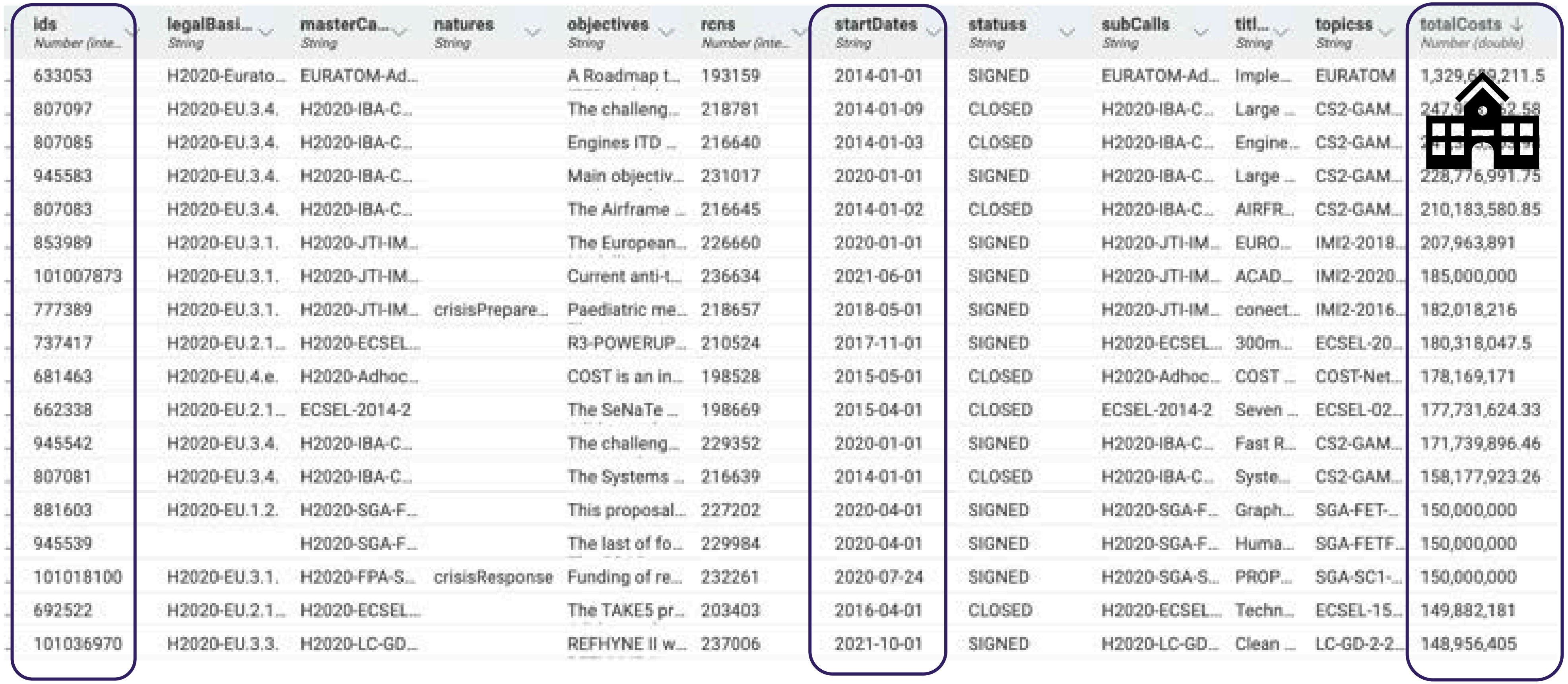


RowID	projectID Number (inte…	projectAc… String	organisat… Number (inte…	vatNumber String	name String	shortName String	SME String	activityTy… String	street String	postCode String	city String	country String
Row0	879926	EEN SACHSEN	999913831	?	INDUSTRIE-U…	IHK C	false	PUB	STRASSE DE…	09111	CHEMNITZ	DE
Row1	879926	EEN SACHSEN	994556036	DE176093942	ZTS-ZENTRU…	ZTS	true	OTH	INDUSTRIEST…	01612	GLAUBITZ	DE
Row2	879926	EEN SACHSEN	999797625	?	INDUSTRIE- U…	IHK DRESDEN	false	PUB	LANGER WEG…	01239	DRESDEN	DE
Row3	879926	EEN SACHSEN	999940409	?	[illegible]	HWK DRESDE…	false	PUB	AM LAGERPL…	[illegible]	DRESDEN	DE
Row4	879926	EEN SACHSEN	999877844	DE140857609	TECHNISCHE…	?	false	HES	STRASSE DE…	[illegible]	Chemnitz	DE
Row5	879926	EEN SACHSEN	999914219	?	INDUSTRIE-U…	IIHK-L	false	PUB	GOERDELERR…	04109	LEIPZIG	DE
Row6	879926	EEN SACHSEN	969246408	?	HOCHSCHUL…	Hochschule Z…	false	HES	THEODOR-KO…	02763	Zittau	DE
Row7	879926	EEN SACHSEN	999917905	DE141484057	AGENTUR FU…	AGIL	false	PRC	LESSINGSTR…	04109	LEIPZIG	DE
Row8	740689	HEIMDALL	997710476	IT015032900…	CENTRO INT…	FONDAZIONE…	false	REC	VIA ARMAND…	17100	SAVONA	IT
Row9	740689	HEIMDALL	940057847	DK37222135	FREDERIKSB…	FBBR	false	PRC	LOGISMOSE 3	3600	FREDERIKSS…	DK
Row…	740689	HEIMDALL	999703438	ESS0811001G	Departament …	INT	false	PUB	Carrer Diputa…	08009	Barcelona	ES
Row…	740689	HEIMDALL	919049102	IT136697210…	ASSOCIAZIO…	?	false	OTH	VIA BERNAR…	00151	Roma	IT
Row…	740689	HEIMDALL	999778322	EL094149709	SPACE HELL…	SPACE HELL…	true	PRC	MESSOGION …	153 41	Aghia Parask…	EL
Row…	740689	HEIMDALL	999638739	ESG62616586	CENTRE TEC…	CTTC	false	REC	AVINGUDA C…	08860	Castelldefels …	ES
Row…	740689	HEIMDALL	996569950	FR44130005…	UNIVERSITE …	UNISTRA	false	HES	RUE BLAISE P…	67081	Strasbourg	FR
Row…	740689	HEIMDALL	935977542	ESQ0801980D	INSTITUT CA…	ICGC	false	PUB	PARC DE MO…	08038	Barcelona	ES
Row…	740689	HEIMDALL	997822026	ESB24352296	TECNOSYLVA…	?	true	PRC	Parque Tecno…	24009	Leon	ES
Row…	740689	HEIMDALL	999981731	DE121965658	DEUTSCHES …	DLR	false	REC	LINDER HOHE	51147	Koln	DE
Row…	740689	HEIMDALL	917614763	GB158535583	SCOTTISH FI…	SCOTTISH FI…	false	PUB	SCOTTISH FI…	G72 7NA	CAMBUSLANG	UK
Row…	740689	HEIMDALL	989367991	GB830703457	AVANTI COM…	?	true	PRC	COBHAM HO…	EC4V 6EB	London	UK
Row…	740689	HEIMDALL	958896217	ESG55536098	FUNDACIO D'…	PAU COSTA F…	true	REC	AV. MOSSEN …	08552	Taradell	ES

- https://data.europa.eu/data/datasets/cordish2020projects?locale=en
- https://data.europa.eu/data/datasets/cordis-eu-research-projects-under-horizon-europe-2021-2027?locale=en

Step 2 - **Identify** columns

- Horizon Data "Projects" - *Identify Energy **Projects** per Country per **year***



ids Number (inte...	legalBasi... String	masterCa... String	natures String	objectives String	rcns Number (inte...	startDates String	statuss String	subCalls String	titl... String	topicss String	totalCosts ↓ Number (double)
633053	H2020-Eurato...	EURATOM-Ad...		A Roadmap t...	193159	2014-01-01	SIGNED	EURATOM-Ad...	Imple...	EURATOM	1,329,6[illegible],211.5
807097	H2020-EU.3.4.	H2020-IBA-C...		The challeng...	218781	2014-01-09	CLOSED	H2020-IBA-C...	Large ...	CS2-GAM...	247,9[illegible]2.58
807085	H2020-EU.3.4.	H2020-IBA-C...		Engines ITD ...	216640	2014-01-03	CLOSED	H2020-IBA-C...	Engine...	CS2-GAM...	[illegible]
945583	H2020-EU.3.4.	H2020-IBA-C...		Main objectiv...	231017	2020-01-01	SIGNED	H2020-IBA-C...	Large ...	CS2-GAM...	228,776,991.75
807083	H2020-EU.3.4.	H2020-IBA-C...		The Airframe ...	216645	2014-01-02	CLOSED	H2020-IBA-C...	AIRFR...	CS2-GAM...	210,183,580.85
853989	H2020-EU.3.1.	H2020-JTI-IM...		The European...	226660	2020-01-01	SIGNED	H2020-JTI-IM...	EURO...	IMI2-2018...	207,963,891
101007873	H2020-EU.3.1.	H2020-JTI-IM...		Current anti-t...	236634	2021-06-01	SIGNED	H2020-JTI-IM...	ACAD...	IMI2-2020...	185,000,000
777389	H2020-EU.3.1.	H2020-JTI-IM...	crisisPrepare...	Paediatric me...	218657	2018-05-01	SIGNED	H2020-JTI-IM...	conect...	IMI2-2016...	182,018,216
737417	H2020-EU.2.1...	H2020-ECSEL...		R3-POWERUP...	210524	2017-11-01	SIGNED	H2020-ECSEL...	300m...	ECSEL-20...	180,318,047.5
681463	H2020-EU.4.e.	H2020-Adhoc...		COST is an in...	198528	2015-05-01	CLOSED	H2020-Adhoc...	COST ...	COST-Net...	178,169,171
662338	H2020-EU.2.1...	ECSEL-2014-2		The SeNaTe ...	198669	2015-04-01	CLOSED	ECSEL-2014-2	Seven ...	ECSEL-02...	177,731,624.33
945542	H2020-EU.3.4.	H2020-IBA-C...		The challeng...	229352	2020-01-01	SIGNED	H2020-IBA-C...	Fast R...	CS2-GAM...	171,739,896.46
807081	H2020-EU.3.4.	H2020-IBA-C...		The Systems ...	216639	2014-01-01	CLOSED	H2020-IBA-C...	Syste...	CS2-GAM...	158,177,923.26
881603	H2020-EU.1.2.	H2020-SGA-F...		This proposal...	227202	2020-04-01	SIGNED	H2020-SGA-F...	Graph...	SGA-FET-...	150,000,000
945539		H2020-SGA-F...		The last of fo...	229984	2020-04-01	SIGNED	H2020-SGA-F...	Huma...	SGA-FETF...	150,000,000
101018100	H2020-EU.3.1.	H2020-FPA-S...	crisisResponse	Funding of re...	232261	2020-07-24	SIGNED	H2020-SGA-S...	PROP...	SGA-SC1-...	150,000,000
692522	H2020-EU.2.1...	H2020-ECSEL...		The TAKE5 pr...	203403	2016-04-01	CLOSED	H2020-ECSEL...	Techn...	ECSEL-15...	149,882,181
101036970	H2020-EU.3.3.	H2020-LC-GD...		REFHYNE II w...	237006	2021-10-01	SIGNED	H2020-LC-GD...	Clean ...	LC-GD-2-2...	148,956,405

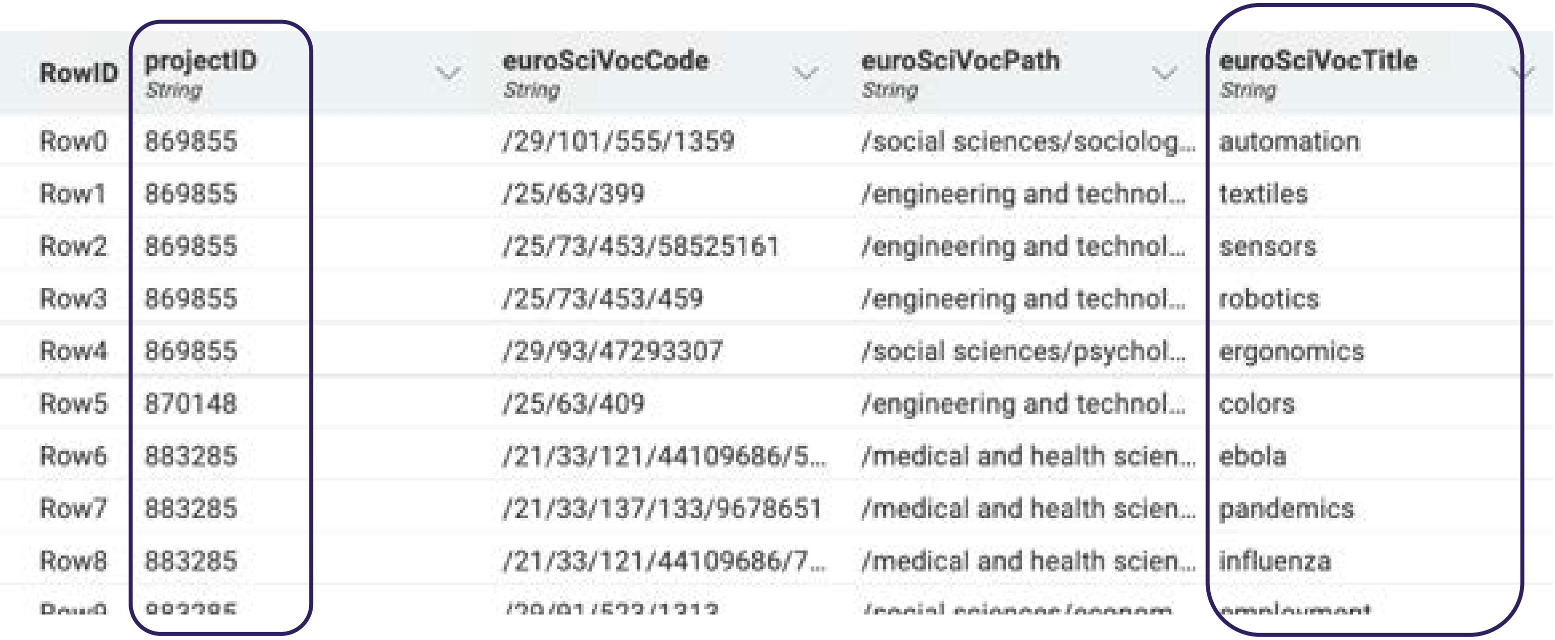
- https://data.europa.eu/data/datasets/cordish2020projects?locale=en
- https://data.europa.eu/data/datasets/cordis-eu-research-projects-under-horizon-europe-2021-2027?locale=en

Step 2 – **Identify** columns

- Horizon Data "EuroSciVoc" - *Identify **Energy** Projects per Country per year*

https://op.europa.eu/en/web/eu-vocabularies/euroscivoc

- Reference vocabulary for the Open Science community and is aligned with Linked Open Data standards.

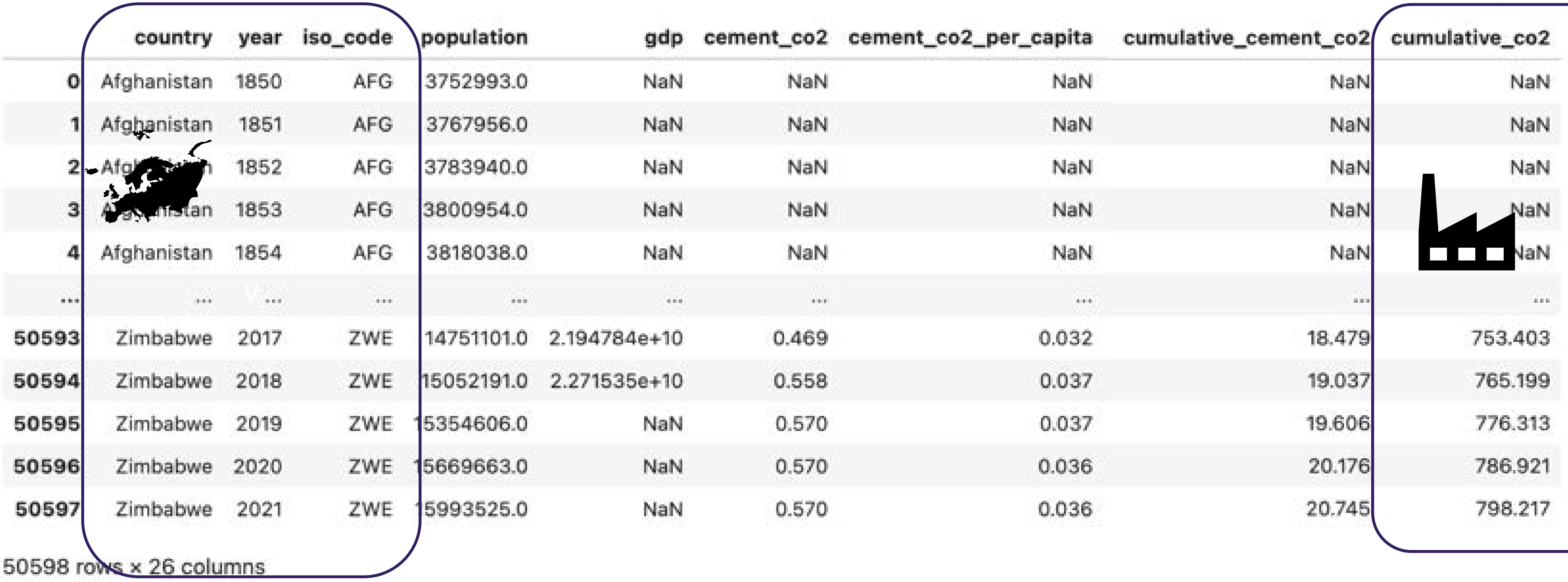


RowID	projectID String	euroSciVocCode String	euroSciVocPath String	euroSciVocTitle String
Row0	869855	/29/101/555/1359	/social sciences/sociolog...	automation
Row1	869855	/25/63/399	/engineering and technol...	textiles
Row2	869855	/25/73/453/58525161	/engineering and technol...	sensors
Row3	869855	/25/73/453/459	/engineering and technol...	robotics
Row4	869855	/29/93/47293307	/social sciences/psychol...	ergonomics
Row5	870148	/25/63/409	/engineering and technol...	colors
Row6	883285	/21/33/121/44109686/5...	/medical and health scien...	ebola
Row7	883285	/21/33/137/133/9678651	/medical and health scien...	pandemics
Row8	883285	/21/33/121/44109686/7...	/medical and health scien...	influenza
Row9	883285	/29/91/523/1313	/social sciences/econom...	employment

- https://data.europa.eu/data/datasets/cordish2020projects?locale=en
- https://data.europa.eu/data/datasets/cordis-eu-research-projects-under-horizon-europe-2021-2027?locale=en

Step 2 – Identify **columns** and **rows**

- CO2 emissions data: *Identify and Visualise **EU Countries** with the Most and Least **CO2 Emissions***



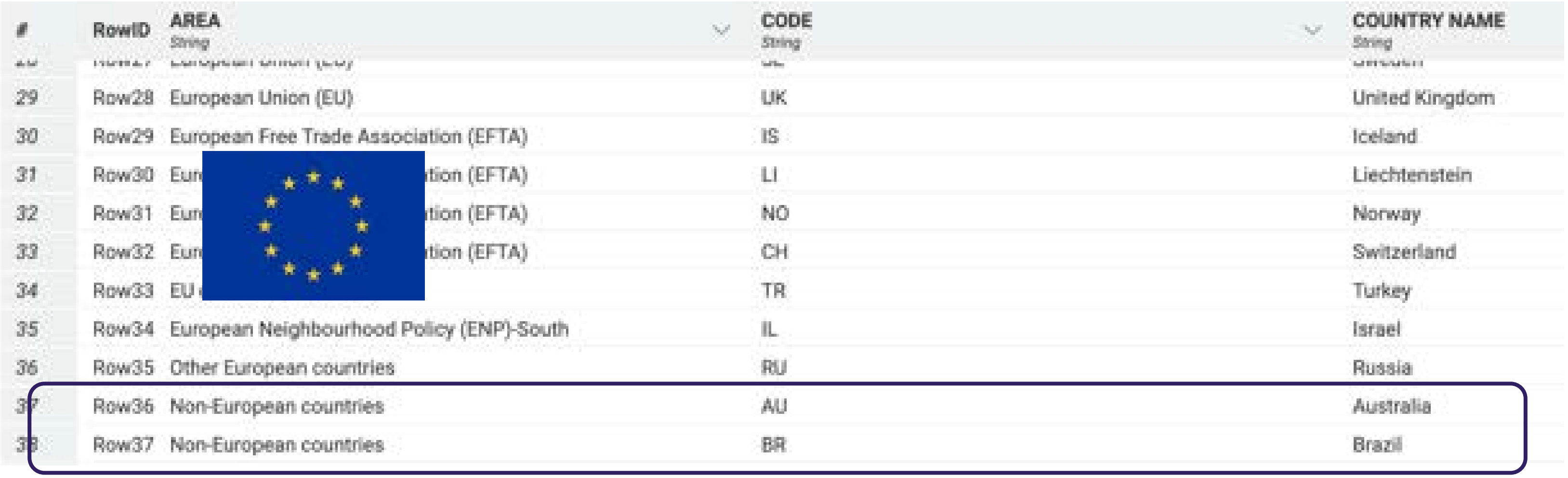
	country	year	iso_code	population	gdp	cement_co2	cement_co2_per_capita	cumulative_cement_co2	cumulative_co2
0	Afghanistan	1850	AFG	3752993.0	NaN	NaN	NaN	NaN	NaN
1	Afghanistan	1851	AFG	3767956.0	NaN	NaN	NaN	NaN	NaN
2	Afghanistan	1852	AFG	3783940.0	NaN	NaN	NaN	NaN	NaN
3	Afghanistan	1853	AFG	3800954.0	NaN	NaN	NaN	NaN	NaN
4	Afghanistan	1854	AFG	3818038.0	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...
50593	Zimbabwe	2017	ZWE	14751101.0	2.194784e+10	0.469	0.032	18.479	753.403
50594	Zimbabwe	2018	ZWE	15052191.0	2.271535e+10	0.558	0.037	19.037	765.199
50595	Zimbabwe	2019	ZWE	15354606.0	NaN	0.570	0.037	19.606	776.313
50596	Zimbabwe	2020	ZWE	15669663.0	NaN	0.570	0.036	20.176	786.921
50597	Zimbabwe	2021	ZWE	15993525.0	NaN	0.570	0.036	20.745	798.217

50598 rows × 26 columns

- https://ourworldindata.org/co2-emissions

Step 2 – Identify **rows**

- EU Country Names data: Filter EU countries (rows)



#	RowID	AREA String	CODE String	COUNTRY NAME String
29	Row28	European Union (EU)	UK	United Kingdom
30	Row29	European Free Trade Association (EFTA)	IS	Iceland
31	Row30	Eur tion (EFTA)	LI	Liechtenstein
32	Row31	Eur tion (EFTA)	NO	Norway
33	Row32	Eur tion (EFTA)	CH	Switzerland
34	Row33	EU	TR	Turkey
35	Row34	European Neighbourhood Policy (ENP)-South	IL	Israel
36	Row35	Other European countries	RU	Russia
37	Row36	Non-European countries	AU	Australia
38	Row37	Non-European countries	BR	Brazil

https://ec.europa.eu/eurostat/statistics-explained/images/9/9f/Country_Codes_and_Names.xlsx

Data Cleaning

Step 1: Explore data

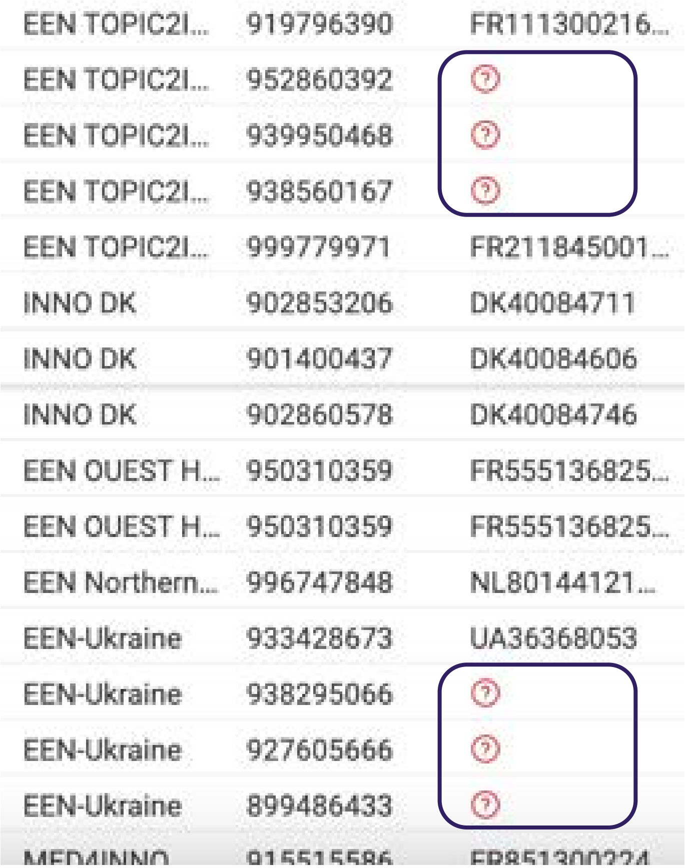
Step 2: Identify which columns/ rows you need, and **filter or sort**

Step 3: Identify if there are missing values or duplicates you need to deal with

Step 3 - Missing Values

- Define how to handle missing values for all columns of a given type
  - Affect all columns that are not explicitly mentioned in the second tab
- Define how to handle missing values for each available column

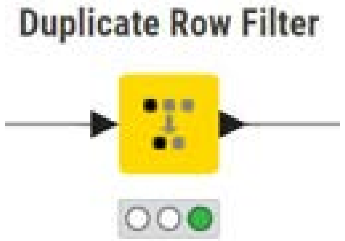
Missing Value
?



EEN TOPIC2I...	919796390	FR111300216...
EEN TOPIC2I...	952860392	?
EEN TOPIC2I...	939950468	?
EEN TOPIC2I...	938560167	?
EEN TOPIC2I...	999779971	FR211845001...
INNO DK	902853206	DK40084711
INNO DK	901400437	DK40084606
INNO DK	902860578	DK40084746
EEN OUEST H...	950310359	FR555136825...
EEN OUEST H...	950310359	FR555136825...
EEN Northern...	996747848	NL80144121...
EEN-Ukraine	933428673	UA36368053
EEN-Ukraine	938295066	?
EEN-Ukraine	927605666	?
EEN-Ukraine	899486433	?
MED4INNO	915515586	FR851300224

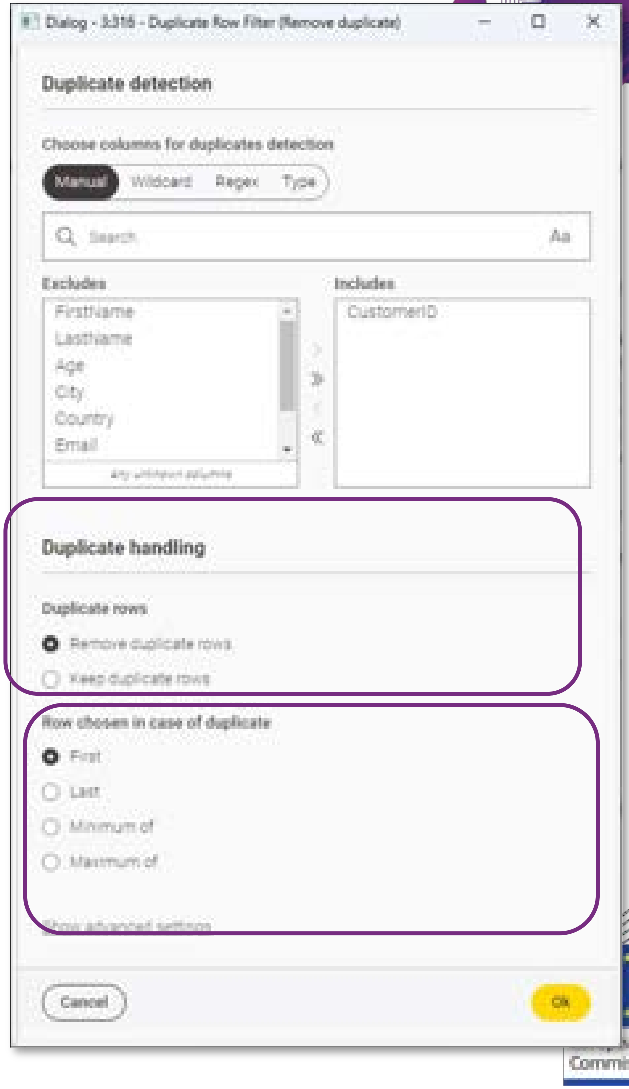
Step 3 - Duplicates values

- Detect duplicate rows and apply a selected treatment
  - Select columns to check for duplicates
  - Provide options for treating duplicated values



Flag or Remove Duplicates

Select criteria to keep row



Dialog - 3:316 - Duplicate Row Filter (Remove duplicate)

Duplicate detection

Choose columns for duplicates detection

Manual | Wildcard | Regex | Type

Search

Excludes	Includes
FirstName	CustomerID
LastName	
Age	
City	
Country	
Email	
Any unknown columns	

Duplicate handling

Duplicate rows

● Remove duplicate rows

○ Keep duplicate rows

Row chosen in case of duplicate

● First

○ Last

○ Minimum of

○ Maximum of

Show advanced settings

Cancel | Ok

Data Cleaning: Recap

Step 1: Explore data

Step 2: Filter or sort

Step 3: Missing values and duplicates

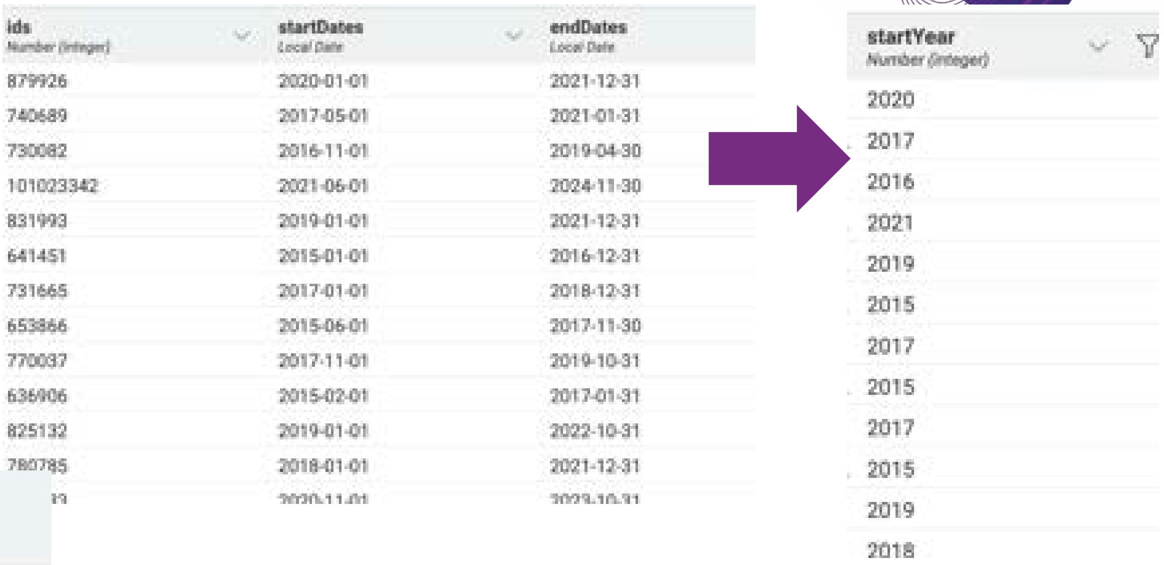
Item	Cost	Store Type	City	VAT	Delivery code
Table	50	Onsite	NY	14%	XY
Couch	100	Online	N-Y	14%	XX
Couch	100	Online	N-Y	14%	XX
Cancelled	?	?	?	?	?
TV	70	Online	NY	14%	XY
Chair	30	Onsite	NY	14%	XY
Table	?	Online	NY	14%	XX

Item	Cost	Store Type	City
Table	50	Onsite	NY
Couch	100	Online	NY
Couch	100	Online	NY
TV	70	Online	NY
Chair	30	Onsite	NY
Table	50	Online	NY

Data Transformation

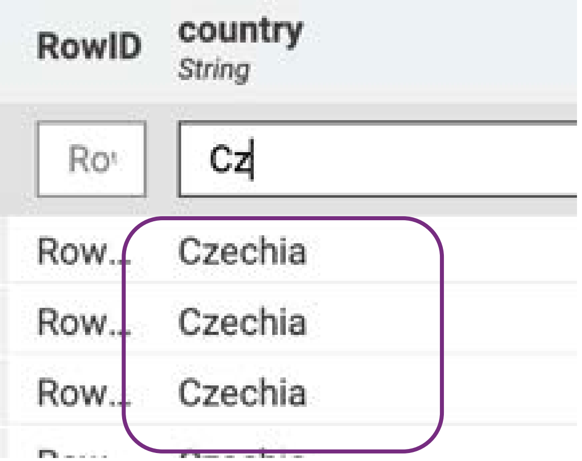
Data Transformation

- **What** do we transform?
  - Shape of the table (new columns)
  - Content of cells

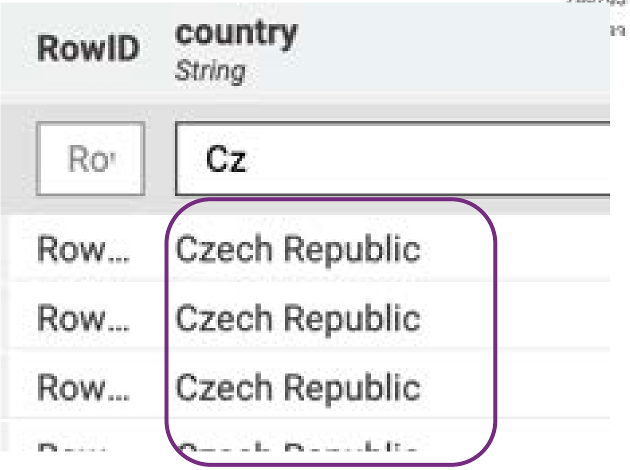


ids *Number (integer)*	startDates *Local Date*	endDates *Local Date*
879926	2020-01-01	2021-12-31
740689	2017-05-01	2021-01-31
730082	2016-11-01	2019-04-30
101023342	2021-06-01	2024-11-30
831993	2019-01-01	2021-12-31
641451	2015-01-01	2016-12-31
731665	2017-01-01	2018-12-31
653866	2015-06-01	2017-11-30
770037	2017-11-01	2019-10-31
636906	2015-02-01	2017-01-31
825132	2019-01-01	2022-10-31
780785	2018-01-01	2021-12-31
[illegible]	2020-11-01	2023-10-31

startYear *Number (integer)*
2020
2017
2016
2021
2019
2015
2017
2015
2017
2015
2019
2018



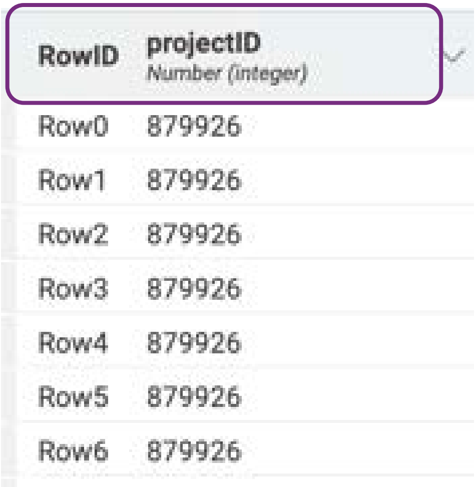
RowID	country *String*
Ro'	Cz
Row...	Czechia
Row...	Czechia
Row...	Czechia



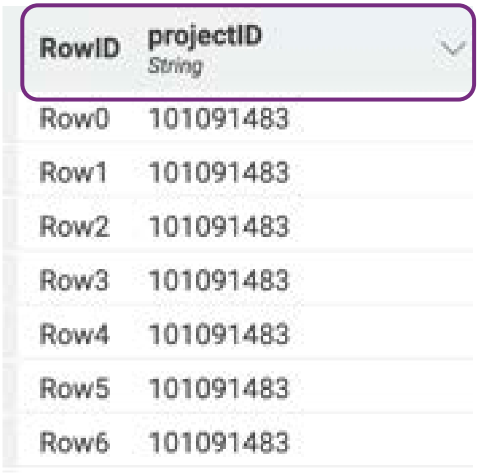
RowID	country *String*
Ro'	Cz
Row...	Czech Republic
Row...	Czech Republic
Row...	Czech Republic

Data Transformation

- **Why transform data?**
  - Columns in mismatching data type that can lead to problems
  - *What are data types?*
    1.**String**: Text or characters.
    2. **Number**: Numerical values.
    3. **Date & Time**: Points in time or durations.
  - Horizon 2020 column "ID" = number
  - Horizon 2021 column "ID" = string



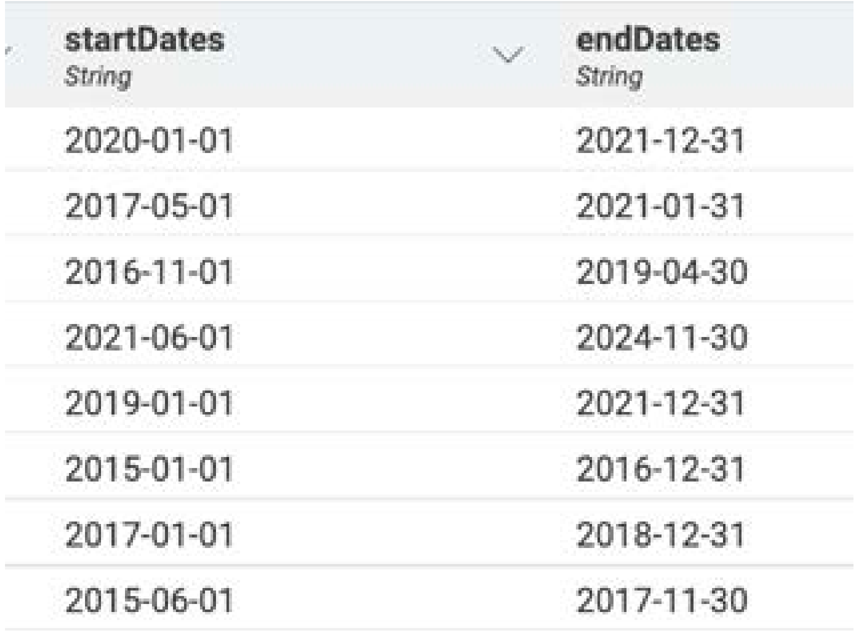
RowID	projectID *Number (integer)*
Row0	879926
Row1	879926
Row2	879926
Row3	879926
Row4	879926
Row5	879926
Row6	879926



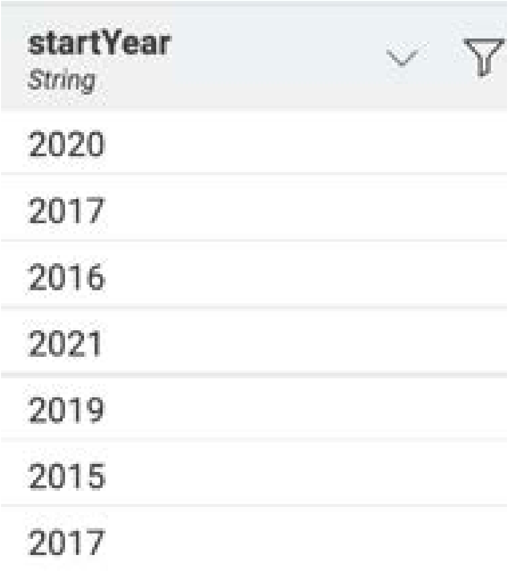
RowID	projectID *String*
Row0	101091483
Row1	101091483
Row2	101091483
Row3	101091483
Row4	101091483
Row5	101091483
Row6	101091483

Data Transformation

- **Why** transform data?
  - Need to extract information that is missing
    - We want the Start Year of the projects
    - "startDate" column, but it comes as a string
    - Need to transform to a date-time format and the extract year!



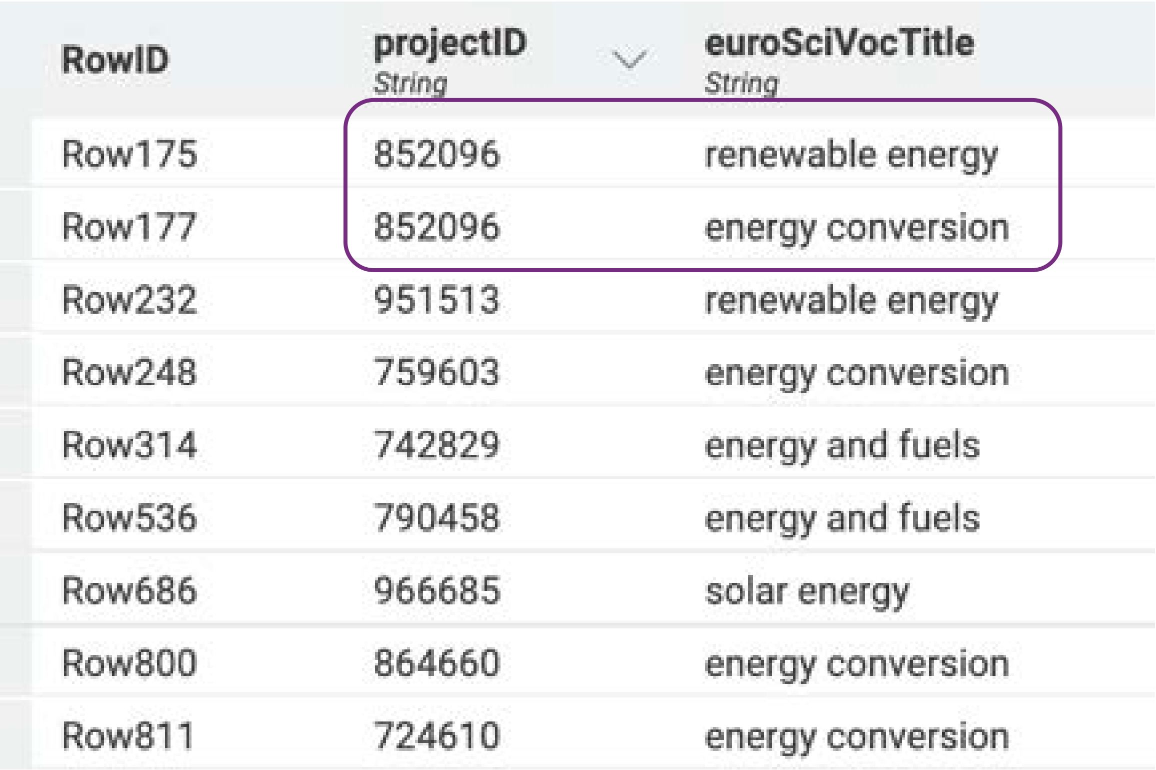
startDates *String*	endDates *String*
2020-01-01	2021-12-31
2017-05-01	2021-01-31
2016-11-01	2019-04-30
2021-06-01	2024-11-30
2019-01-01	2021-12-31
2015-01-01	2016-12-31
2017-01-01	2018-12-31
2015-06-01	2017-11-30



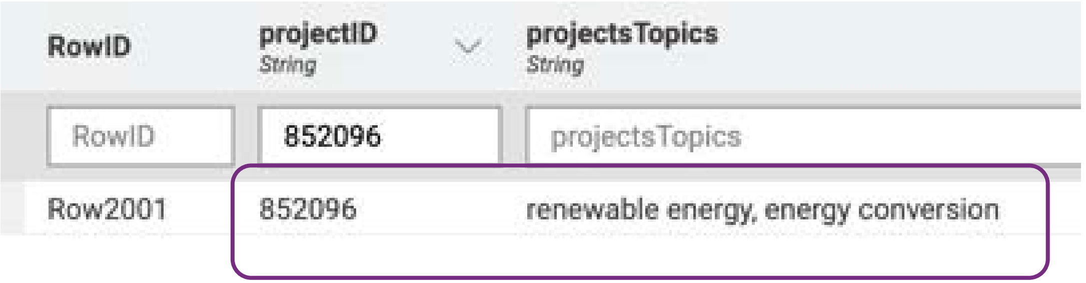
startYear *String*
2020
2017
2016
2021
2019
2015
2017

Data Transformation

- **Why** transform data?
  - One project can have multiple topic name, so we need to **group** the different topics per project



RowID	projectID *String*	euroSciVocTitle *String*
Row175	852096	renewable energy
Row177	852096	energy conversion
Row232	951513	renewable energy
Row248	759603	energy conversion
Row314	742829	energy and fuels
Row536	790458	energy and fuels
Row686	966685	solar energy
Row800	864660	energy conversion
Row811	724610	energy conversion



RowID	projectID *String*	projectsTopics *String*
RowID	852096	projectsTopics
Row2001	852096	renewable energy, energy conversion

Data Transformation

- **Why** transform data?
  - One project can have multiple topic name, so we need to **group** the different topics per project

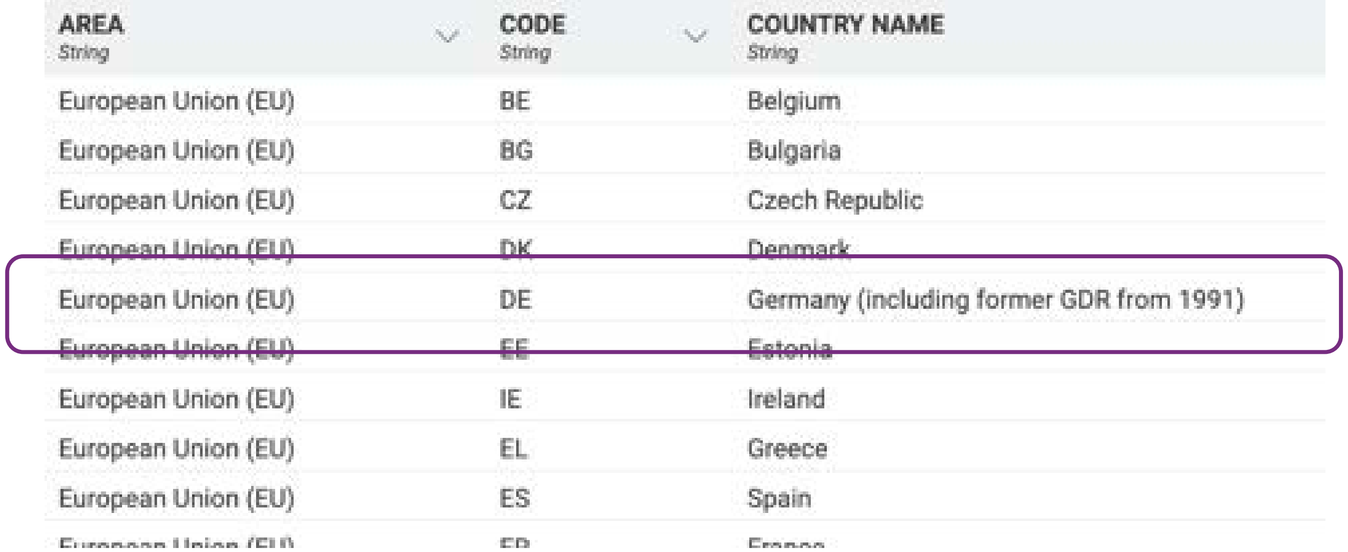
Project ID	Titles
1	Renewable energy
1	Energy conversion
2	solar energy
3	Solar energy
3	Energy conversion
4	Renewable energy

Project ID	Concatenate(Titles)
1	Renewable energy, Energy conversion
2	Solar energy
3	Solar energy, energy conversion
4	Renewable energy

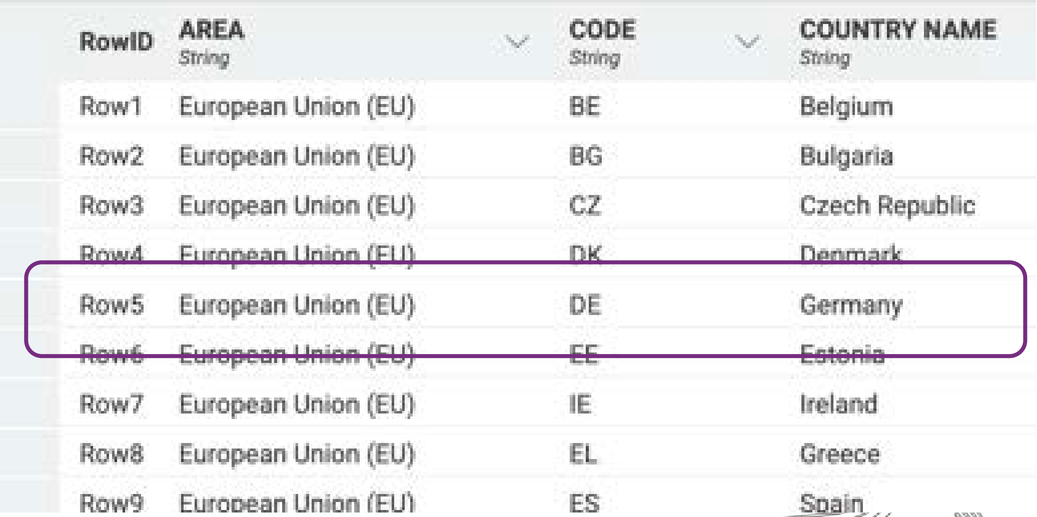
Aggregated on Project ID(group) by Concatenation (aggregation method)

Data Transformation

- **Why** transform data?
  - Some country names are incorrect and can also lead to problems later when merging with all the date
    - Transform string and get them in the correct format



AREA String	CODE String	COUNTRY NAME String
European Union (EU)	BE	Belgium
European Union (EU)	BG	Bulgaria
European Union (EU)	CZ	Czech Republic
European Union (EU)	DK	Denmark
European Union (EU)	DE	Germany (including former GDR from 1991)
European Union (EU)	EE	Estonia
European Union (EU)	IE	Ireland
European Union (EU)	EL	Greece
European Union (EU)	ES	Spain
European Union (EU)	FR	France



RowID	AREA String	CODE String	COUNTRY NAME String
Row1	European Union (EU)	BE	Belgium
Row2	European Union (EU)	BG	Bulgaria
Row3	European Union (EU)	CZ	Czech Republic
Row4	European Union (EU)	DK	Denmark
Row5	European Union (EU)	DE	Germany
Row6	European Union (EU)	EE	Estonia
Row7	European Union (EU)	IE	Ireland
Row8	European Union (EU)	EL	Greece
Row9	European Union (EU)	ES	Spain

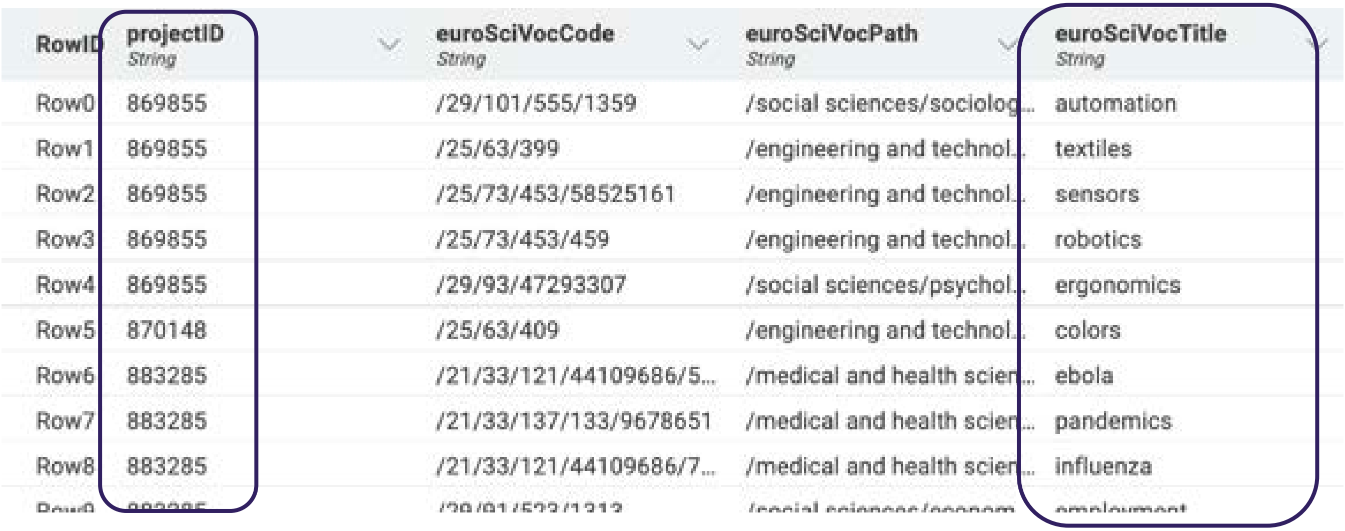
Recap

Data Transformation: **what** and **why**?

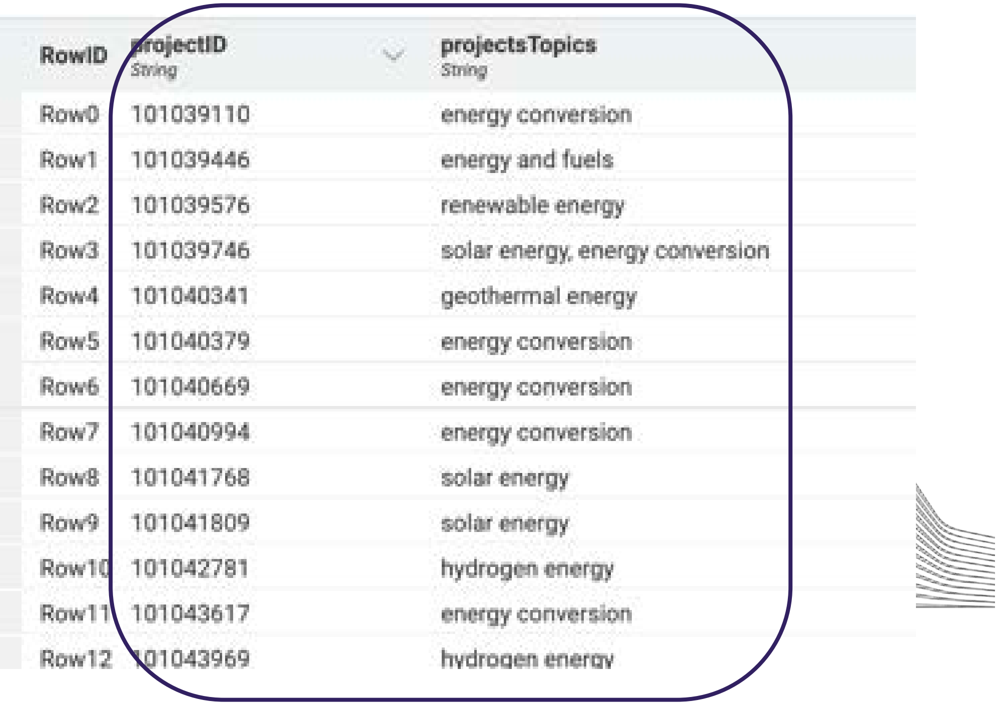
From

raw data

To transformed data in a format tailored to Zoi's needs



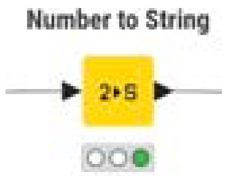
RowID	projectID	euroSciVocCode	euroSciVocPath	euroSciVocTitle
	String	String	String	String
Row0	869855	/29/101/555/1359	/social sciences/sociolog...	automation
Row1	869855	/25/63/399	/engineering and technol...	textiles
Row2	869855	/25/73/453/58525161	/engineering and technol...	sensors
Row3	869855	/25/73/453/459	/engineering and technol...	robotics
Row4	869855	/29/93/47293307	/social sciences/psychol...	ergonomics
Row5	870148	/25/63/409	/engineering and technol...	colors
Row6	883285	/21/33/121/44109686/5...	/medical and health scien...	ebola
Row7	883285	/21/33/137/133/9678651	/medical and health scien...	pandemics
Row8	883285	/21/33/121/44109686/7...	/medical and health scien...	influenza
Row9	883285	/29/91/523/1313	/social sciences/econom...	employment



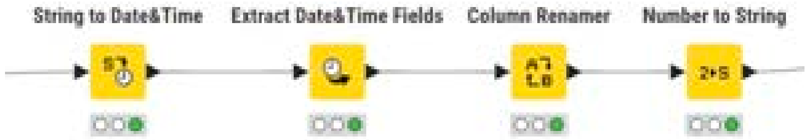
RowID	projectID	projectsTopics
	String	String
Row0	101039110	energy conversion
Row1	101039446	energy and fuels
Row2	101039576	renewable energy
Row3	101039746	solar energy, energy conversion
Row4	101040341	geothermal energy
Row5	101040379	energy conversion
Row6	101040669	energy conversion
Row7	101040994	energy conversion
Row8	101041768	solar energy
Row9	101041809	solar energy
Row10	101042781	hydrogen energy
Row11	101043617	energy conversion
Row12	101043969	hydrogen energy

Data Transformation: How?

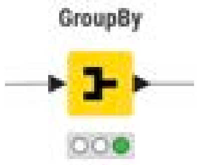
- ID Column in Horizon 2020 to the correct data format



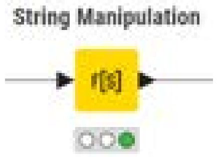
- We extracted the Start Year of projects



- Grouped multiple titles into a single row for each project

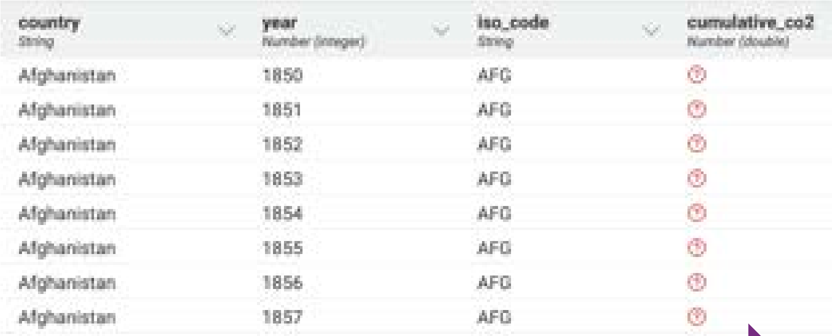


- Gave Countries the correct wording



A bit more cleaning

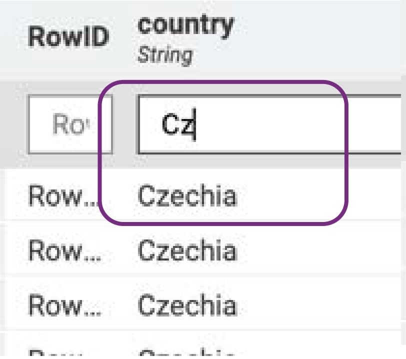
Sometimes we need to clean again after doing some transformations!



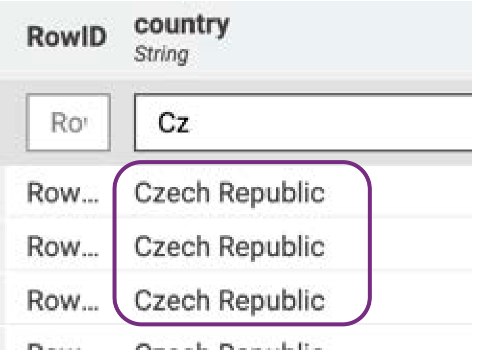
country String	year Number (integer)	iso_code String	cumulative_co2 Number (double)
Afghanistan	1850	AFG	?
Afghanistan	1851	AFG	?
Afghanistan	1852	AFG	?
Afghanistan	1853	AFG	?
Afghanistan	1854	AFG	?
Afghanistan	1855	AFG	?
Afghanistan	1856	AFG	?
Afghanistan	1857	AFG	?

Clean

Transform

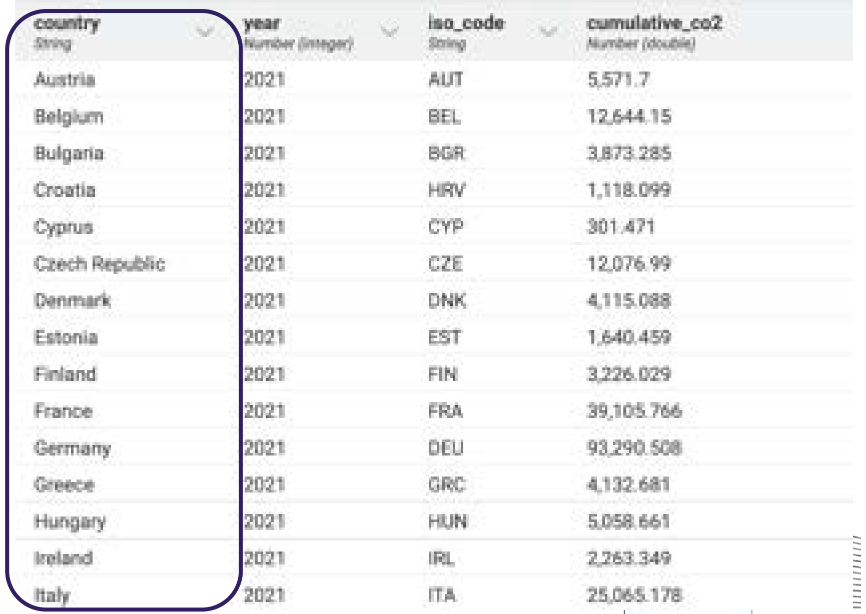


RowID	country String
Ro	Cz
Row...	Czechia
Row...	Czechia
Row...	Czechia



RowID	country String
Ro	Cz
Row...	Czech Republic
Row...	Czech Republic
Row...	Czech Republic

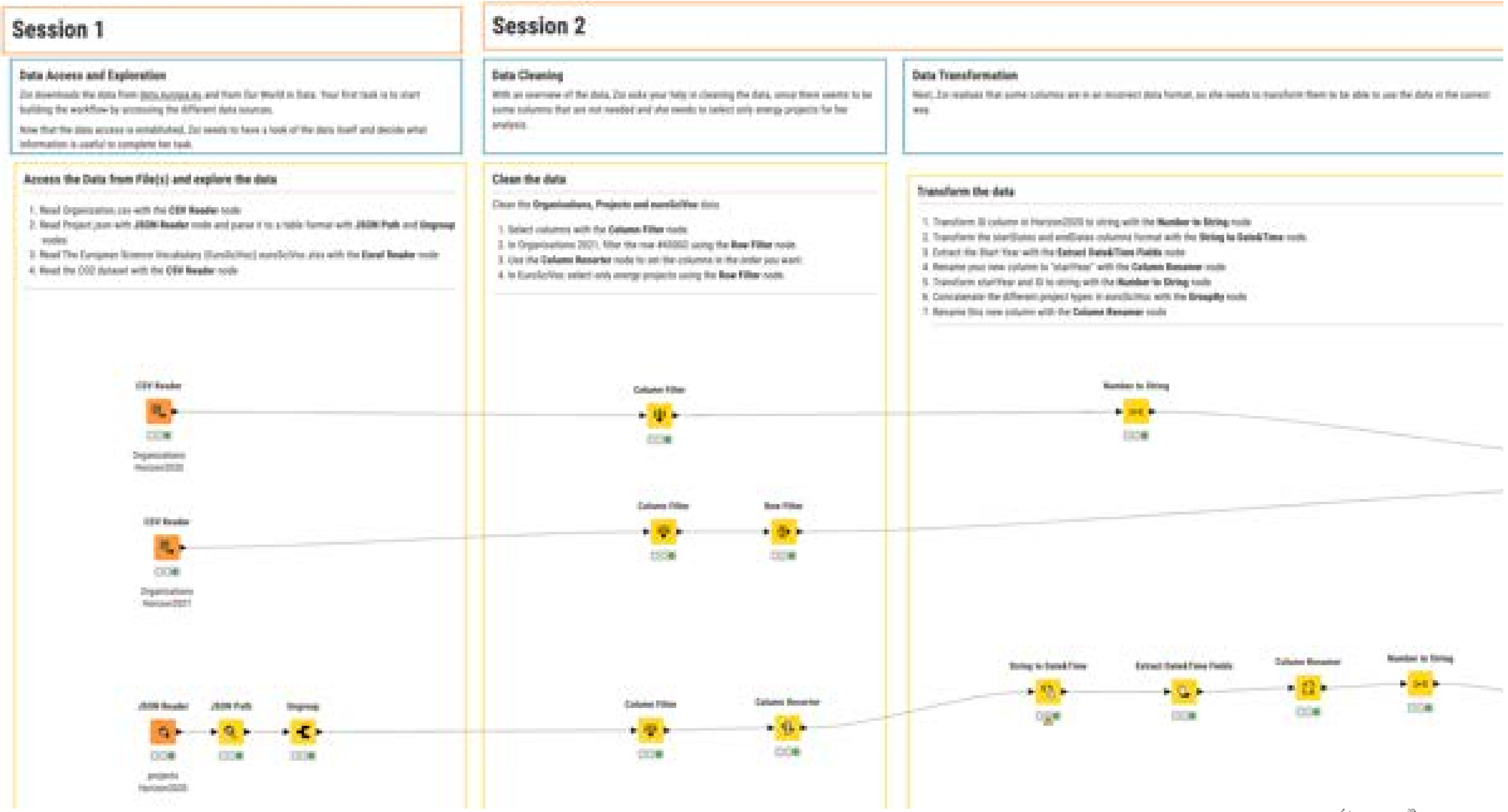
Clean again!



country String	year Number (integer)	iso_code String	cumulative_co2 Number (double)
Austria	2021	AUT	5,571.7
Belgium	2021	BEL	12,644.15
Bulgaria	2021	BGR	3,873.285
Croatia	2021	HRV	1,118.099
Cyprus	2021	CYP	301.471
Czech Republic	2021	CZE	12,076.99
Denmark	2021	DNK	4,115.088
Estonia	2021	EST	1,640.459
Finland	2021	FIN	3,226.029
France	2021	FRA	39,105.766
Germany	2021	DEU	93,290.508
Greece	2021	GRC	4,132.681
Hungary	2021	HUN	5,058.661
Ireland	2021	IRL	2,263.349
Italy	2021	ITA	25,065.178

Today's Output: Workflow

- All these steps are in one single workflow that we can re-run and do it again!



Session 1

Data Access and Exploration

Zoi downloads the data from data.europa.eu and from Our World in Data. Your first task is to start building the workflow by accessing the different data sources.

Now that the data access is established, Zoi needs to have a look of the data itself and decide what information is useful to complete her task.

Access the Data from File(s) and explore the data

1. Read Organization.csv with the **CSV Reader** node
2. Read Project.json with **JSON Reader** node and parse it to a table format with **JSON Path** and **Ungroup** nodes
3. Read The European Science Vocabulary (EuroSciVoc) euroSciVoc.xlsx with the **Excel Reader** node
4. Read the CO2 dataset with the **CSV Reader** node

Session 2

Data Cleaning

With an overview of the data, Zoi asks your help in cleaning the data, since there seems to be some columns that are not needed and she needs to select only energy projects for her analysis.

Clean the data

Clean the **Organizations, Projects and euroSciVoc** data

1. Select columns with the **Column Filter** node
2. In Organizations 2021, filter the row #00000 using the **Row Filter** node
3. Use the **Column Resorter** node to set the columns in the order you want
4. In EuroSciVoc select only energy projects using the **Row Filter** node

Data Transformation

Next, Zoi realises that some columns are in an incorrect data format, so she needs to transform them to be able to use the data in the correct way.

Transform the data

1. Transform ID column in Horizon2020 to string with the **Number to String** node
2. Transform the startDates and endDates columns format with the **String to Date&Time** node
3. Extract the Start Year with the **Extract Date&Time Fields** node
4. Rename your new column to "startYear" with the **Column Renamer** node
5. Transform startYear and ID to string with the **Number to String** node
6. Concatenate the different project types in euroSciVoc with the **GroupBy** node
7. Rename this new column with the **Column Renamer** node

Summary

Now, you should be able to:

- Sign in into the BDTI framework and initialise KNIME

- Prepare the data for analysis by **cleaning** the dataset

- Prepare the data for subsequent analysis by **transforming** the dataset into the correct format

- Address data quality issues

Next session

Data Access

Data Cleaning and Transformation

Data Blending and Export

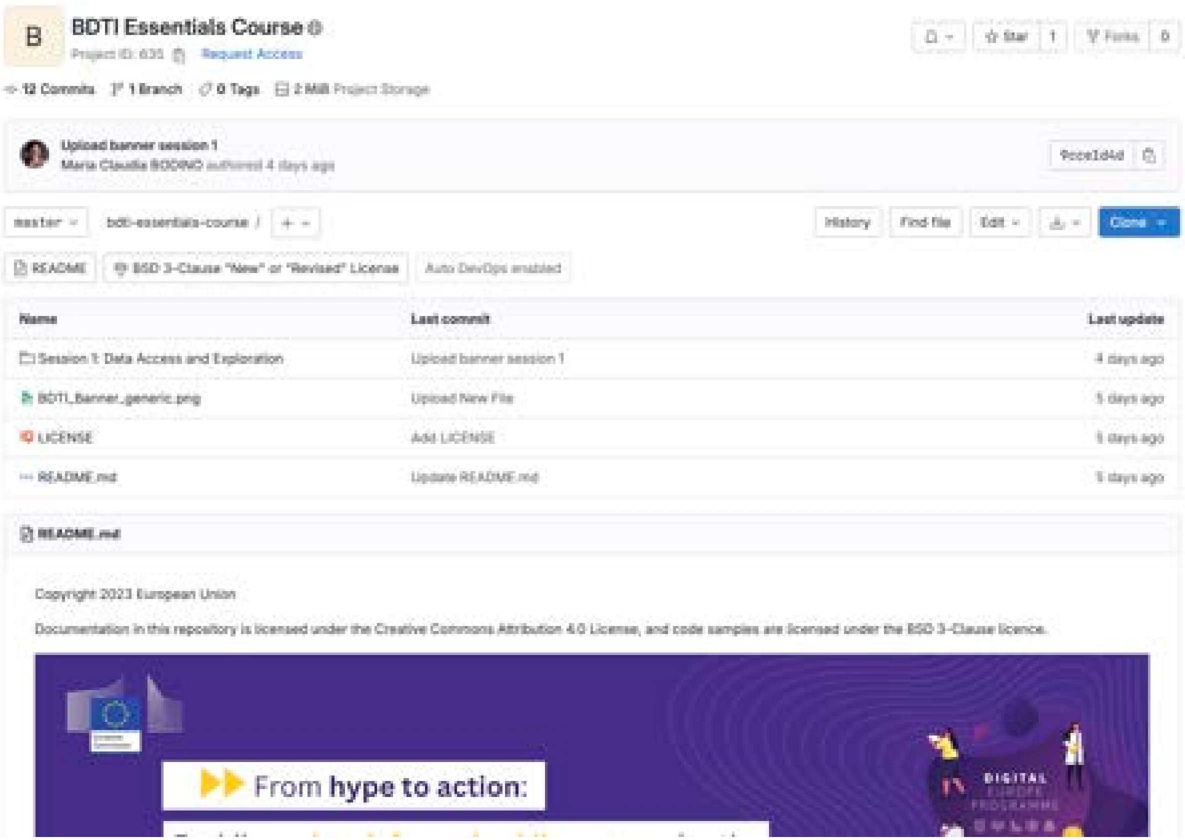
Analytics: aggregation, visualisation, reporting

Advanced Analytics

Practice, practice and practice!

- Slides, hands-on exercises, and solutions will be provided on the webinar Gitlab repository

https://code.europa.eu/bdti/bdti-essentials-course



BDTI Essentials Course

Project ID: 635 Request Access

12 Commits 1 Branch 0 Tags 2 MiB Project Storage

Upload banner session 1
Maria Claudia BODINO authored 4 days ago

9cce1d4d

master bdti-essentials-course /

README BSD 3-Clause "New" or "Revised" License Auto DevOps enabled

Name	Last commit	Last update
Session 1: Data Access and Exploration	Upload banner session 1	4 days ago
BDTI_Banner_generic.png	Upload New File	5 days ago
LICENSE	Add LICENSE	5 days ago
README.md	Update README.md	5 days ago

README.md

Copyright 2023 European Union

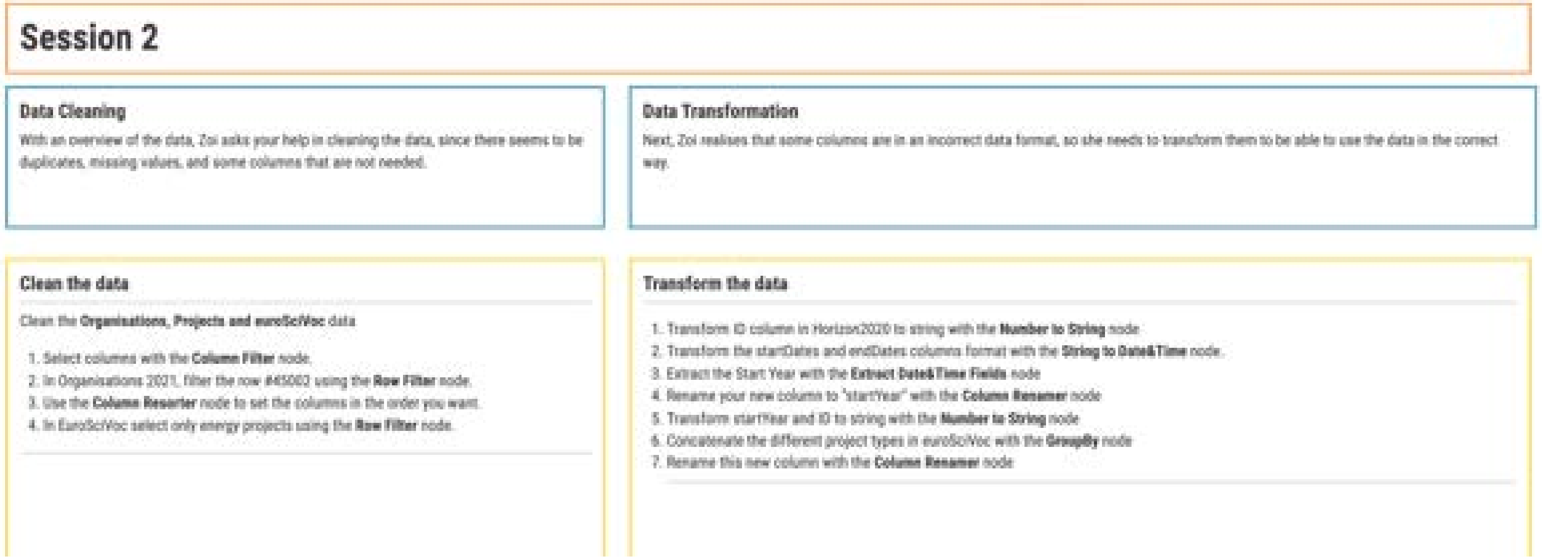
Documentation in this repository is licensed under the Creative Commons Attribution 4.0 License, and code samples are licensed under the BSD 3-Clause licence.

From **hype to action:**

Practice, practice and practice!

- Slides, hands-on exercises, and solutions will be provided on the webinar Gitlab repository

https://code.europa.eu/bdti/bdti-essentials-course



Session 2

Data Cleaning

With an overview of the data, Zoi asks your help in cleaning the data, since there seems to be duplicates, missing values, and some columns that are not needed.

Data Transformation

Next, Zoi realises that some columns are in an incorrect data format, so she needs to transform them to be able to use the data in the correct way.

Clean the data

Clean the **Organisations, Projects and euroSciVoc** data

1. Select columns with the **Column Filter** node.
2. In Organisations 2021, filter the row #45003 using the **Row Filter** node.
3. Use the **Column Resorter** node to set the columns in the order you want.
4. In EuroSciVoc select only energy projects using the **Row Filter** node.

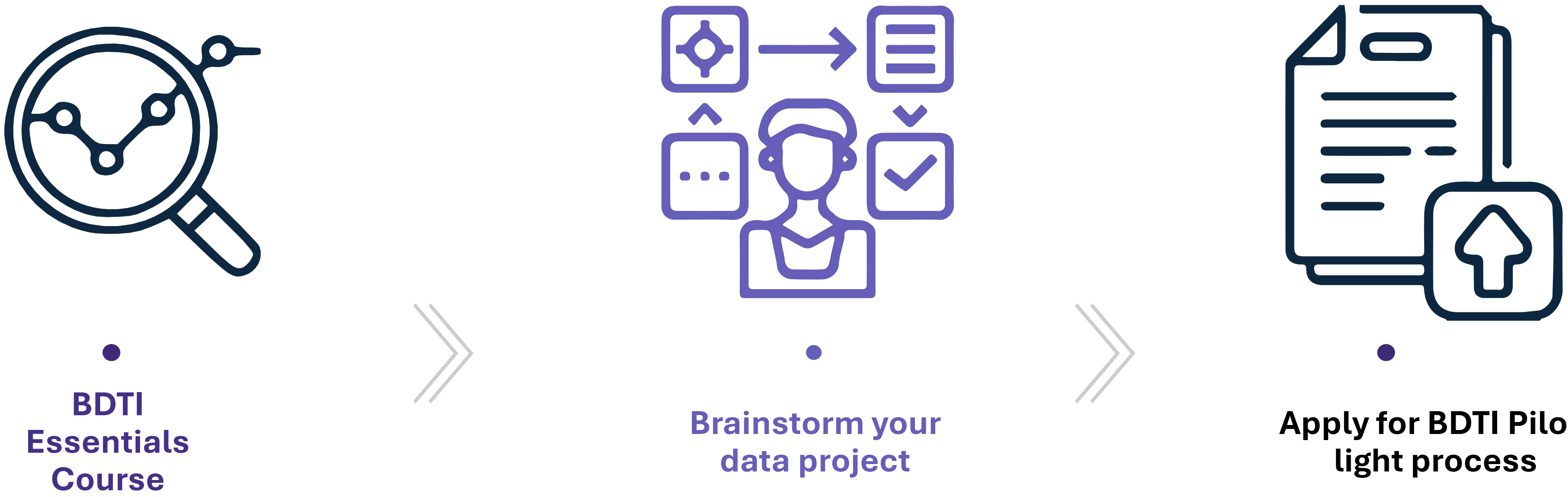
Transform the data

1. Transform ID column in Horizon2020 to string with the **Number to String** node
2. Transform the startDates and endDates columns format with the **String to Date&Time** node.
3. Extract the Start Year with the **Extract Date&Time Fields** node
4. Rename your new column to "startYear" with the **Column Renamer** node
5. Transform startYear and ID to string with the **Number to String** node
6. Concatenate the different project types in euroSciVoc with the **GroupBy** node
7. Rename this new column with the **Column Renamer** node

Q&A

Your BDTI journey starts here

Congrats! You are on the first step to data-driven innovation.

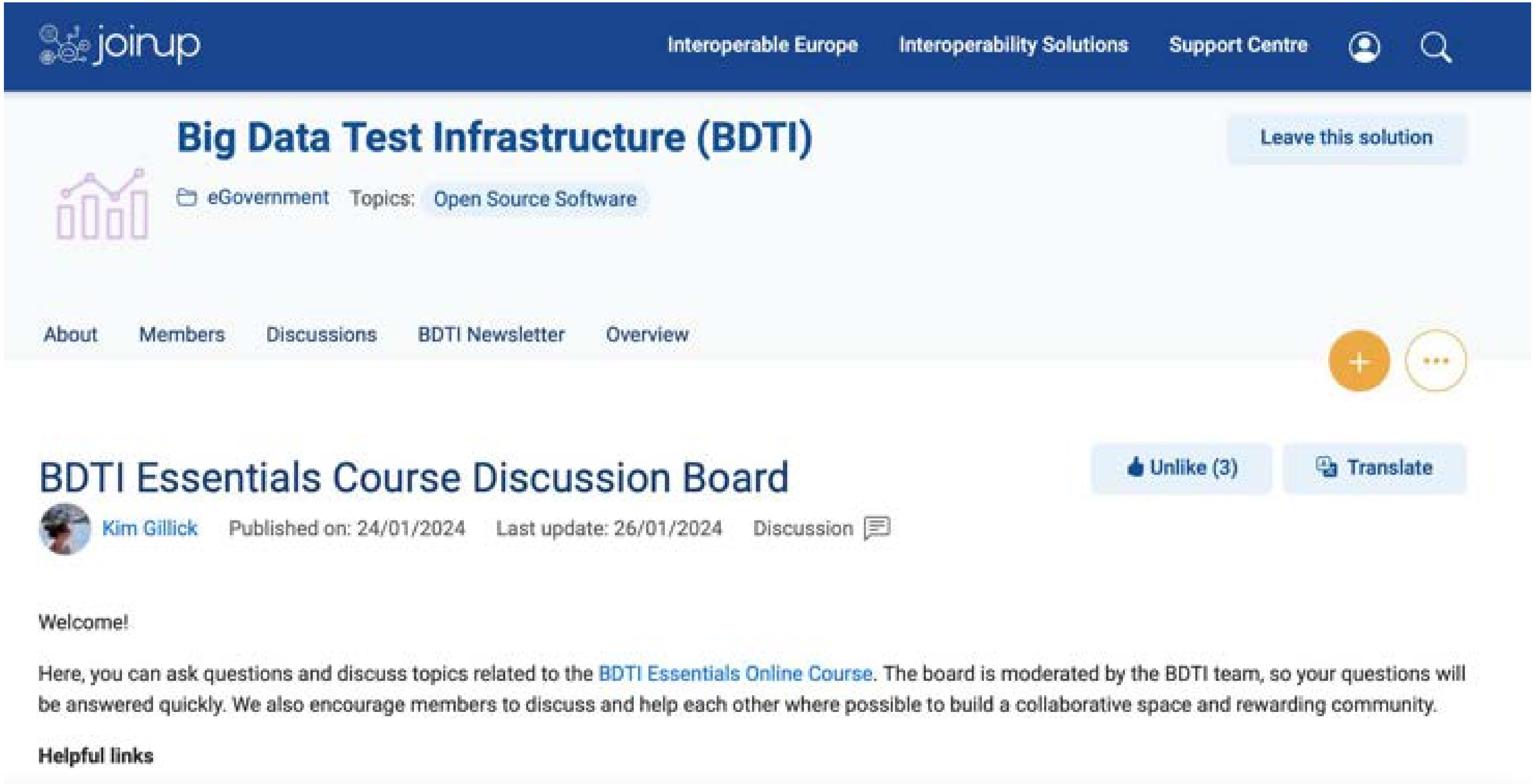


BDTI Essentials Course

Brainstorm your data project

Apply for BDTI Pilot-light process

Course discussion board



Big Data Test Infrastructure (BDTI)

BDTI Essentials Course Discussion Board

Kim Gillick Published on: 24/01/2024 Last update: 26/01/2024 Discussion

Welcome!

Here, you can ask questions and discuss topics related to the BDTI Essentials Online Course. The board is moderated by the BDTI team, so your questions will be answered quickly. We also encourage members to discuss and help each other where possible to build a collaborative space and rewarding community.

Helpful links

Get in touch and follow the BDTI activities

DIGITAL EUROPE PROGRAMME

Are you working for a public administration in need of infrastructure for data analytics?

EC-BDTI-PILOTS@ec.europa.eu

Visit BDTI's website

Subscribe to BDTI's newsletter

Subscribe to BDTI's Joinup

Register for BDTI Essentials online course

https://big-data-test-infrastructure.ec.europa.eu/

European Commission