CENTRAL & EASTERN EUROPEAN SUSTAINABLE ENERGY NETWORK

CEESEN-BENDER

Statistical Tool to Prioritise Buildings for Renovation: User Manual

Co-funded by the European Union

WP4 Task 4.1 Design and test digital tool to prioritise buildings for renovations

CEESEN-BENDER project has developed a statistical tool that uses digital socioeconomic data, technical characteristics of buildings and energy consumption data to generate rankings of buildings that are the least energy-efficient and having occupants most likely to suffer from high levels of energy poverty.

For the analysis, R gramme should be installed.

CENTRAL & EASTERN EUROPEAN SUSTAINABLE ENERGY NETWORK

CEESEN-BENDER

R programme installation

Co-funded by
the European Union

Introduction to R

For statistical modelling of digital tool, which ranks buildings for renovation, R programme has been applied. This software has a number of advantages:

- it is open-source software, meaning it is freely available for anyone to use.
- R runs on various operating systems, including Windows, macOS, and Linux, making it accessible to a broad range of users.
- it has a large and active community of users and developers, providing extensive documentation, forums, and resources for troubleshooting and learning.
- it can easily integrate with other programming languages (like Python, C++, and Java) and tools (like SQL databases and web applications).
- it supports reproducible research through R Markdown, allowing users to create dynamic reports that combine code, results, and narrative text in a single document.
- it is widely used in academia, and there are numerous resources, courses, and textbooks available for learning statistical analysis and R programming.
- R provides a vast array of statistical techniques, including linear and nonlinear modelling, time-series analysis, classification, clustering, and more. It is continuously updated with new statistical methods.

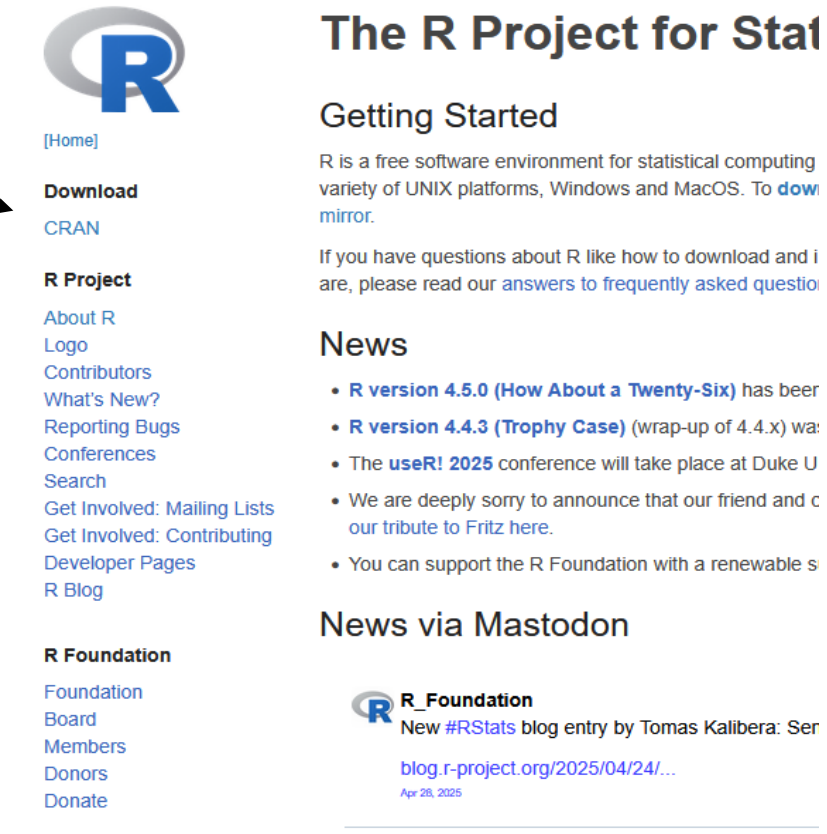
R installation

The latest R can be downloaded from R homepage http://www.r-project.org/:

1. Choose "Download" from the left menu

2. Then choose a mirror server from where to download R (preferably a server closest to you).

3. Specify your operating system (Windows / Mac / Linux) and choose base (base program).

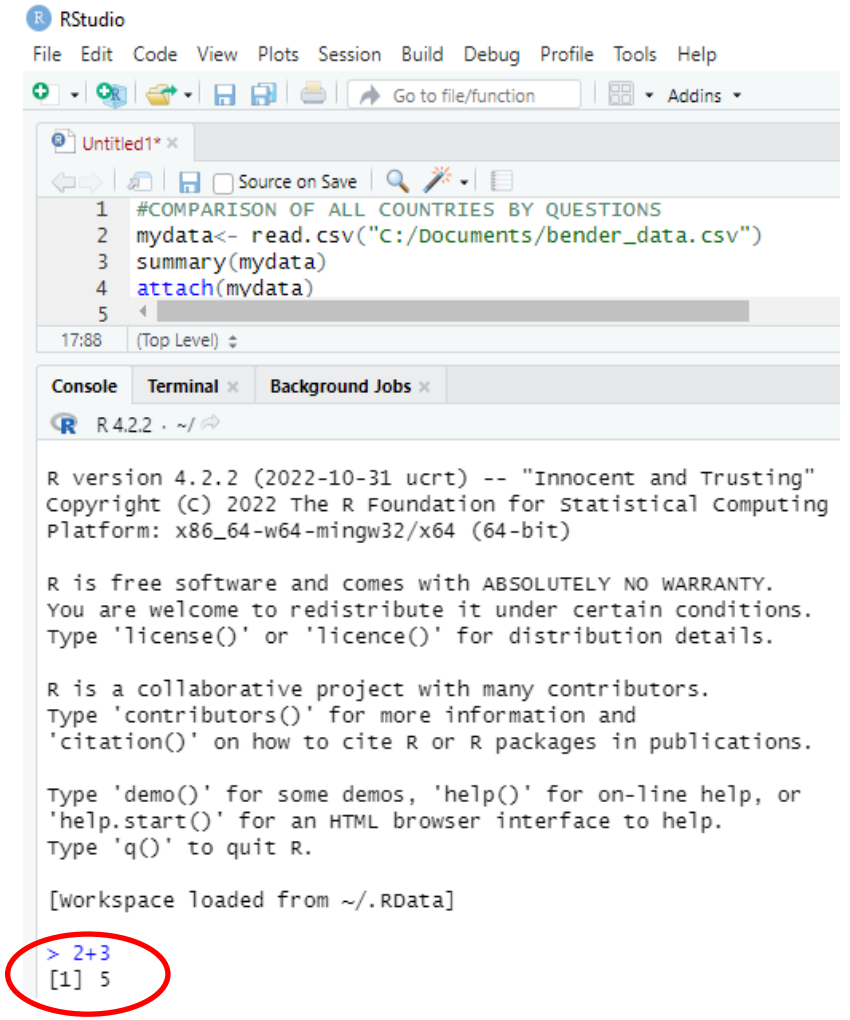


RStudio installation

4. After installing R you can install RStudio which is an IDE (integrated development environment), that makes the R experience much more easier and efficient. For a free version of RStudio go to the following website:

   https://www.rstudio.com/products/rstudio/download/

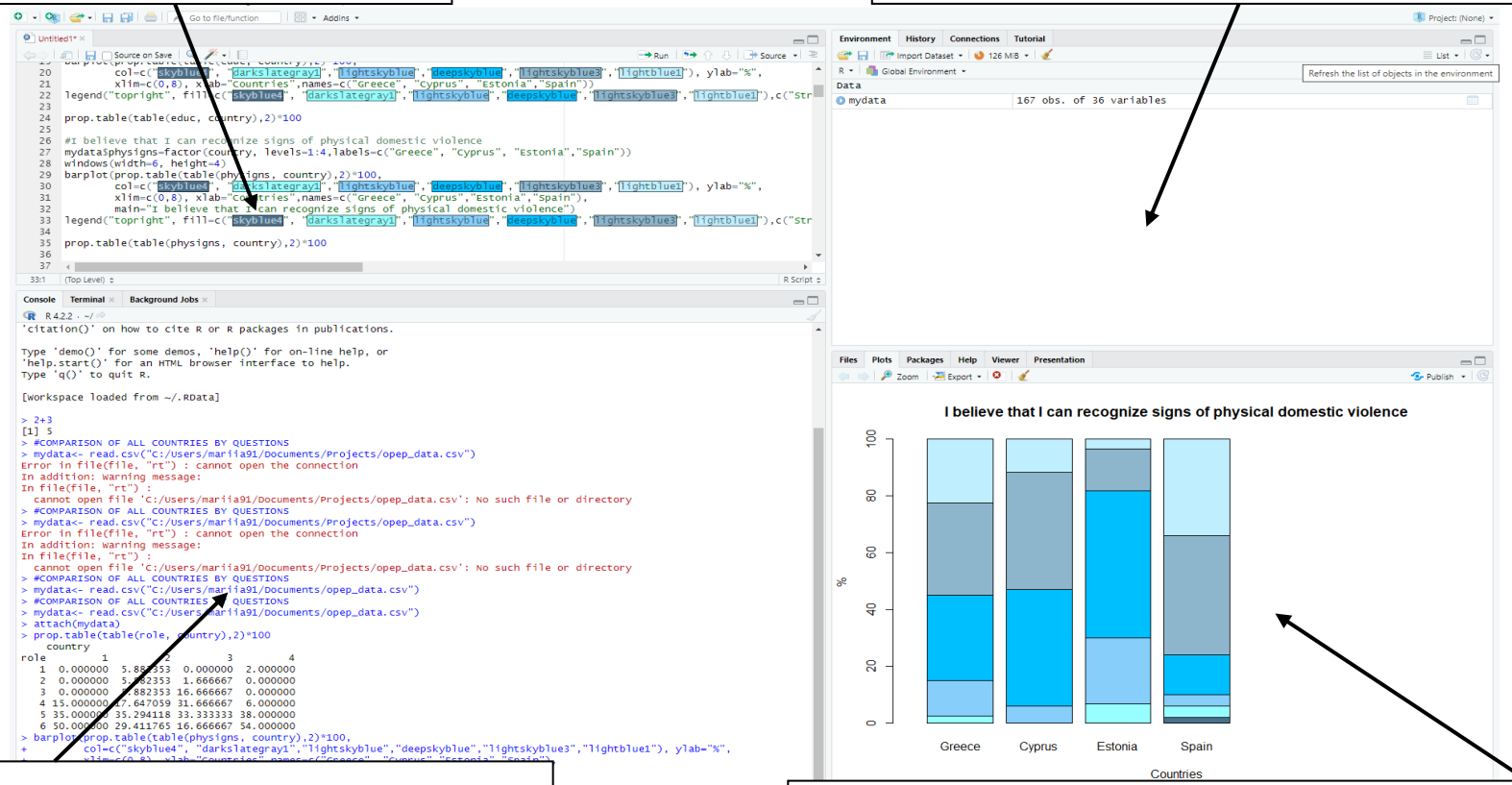
5. Choose the free version for your operating system. Check if all of the installations worked by running RStudio and test if you will get the right answer for 2+3. Combination of Ctrl+Enter executes the command.



R environment

This is the **script window**, the commands that make the analysis are written here. The script can be saved in order to repeat the analysis later.

This is the **environment window**. Here we can see datasets, fitted models and other defined (saved) variables/objects that we can use.



Console window. This is where results, error messages and warnings are shown. You can copy commands straight into this window (if you wish to run a quick calculation not related to the current analysis).

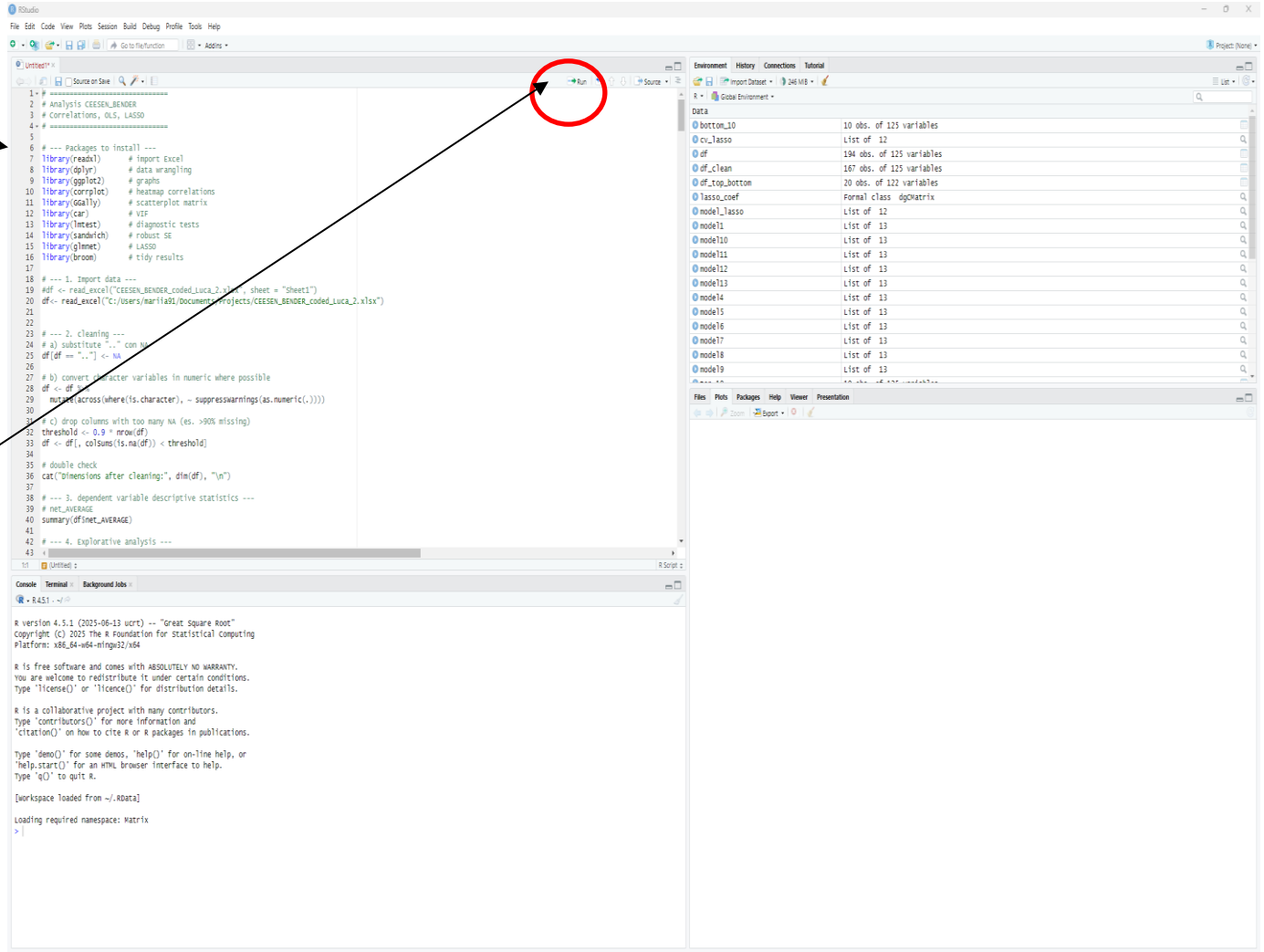
This is where the **plot** will appear. Also, there are tabs on top, one of which is the help tab, containing helpful information on all the functions.

Installation of additional R packages for the analysis (1)

1. Copy these codes and paste into the script window:

```
library(readxl)       # import Excel
library(dplyr)        # data wrangling
library(ggplot2)      # graphs
library(corrplot)     # heatmap correlations
library(GGally)       # scatterplot matrix
library(car)          # VIF
library(lmtest)       # diagnostic tests
library(sandwich)     # robust SE
library(glmnet)       # LASSO
library(broom)        # tidy results
```

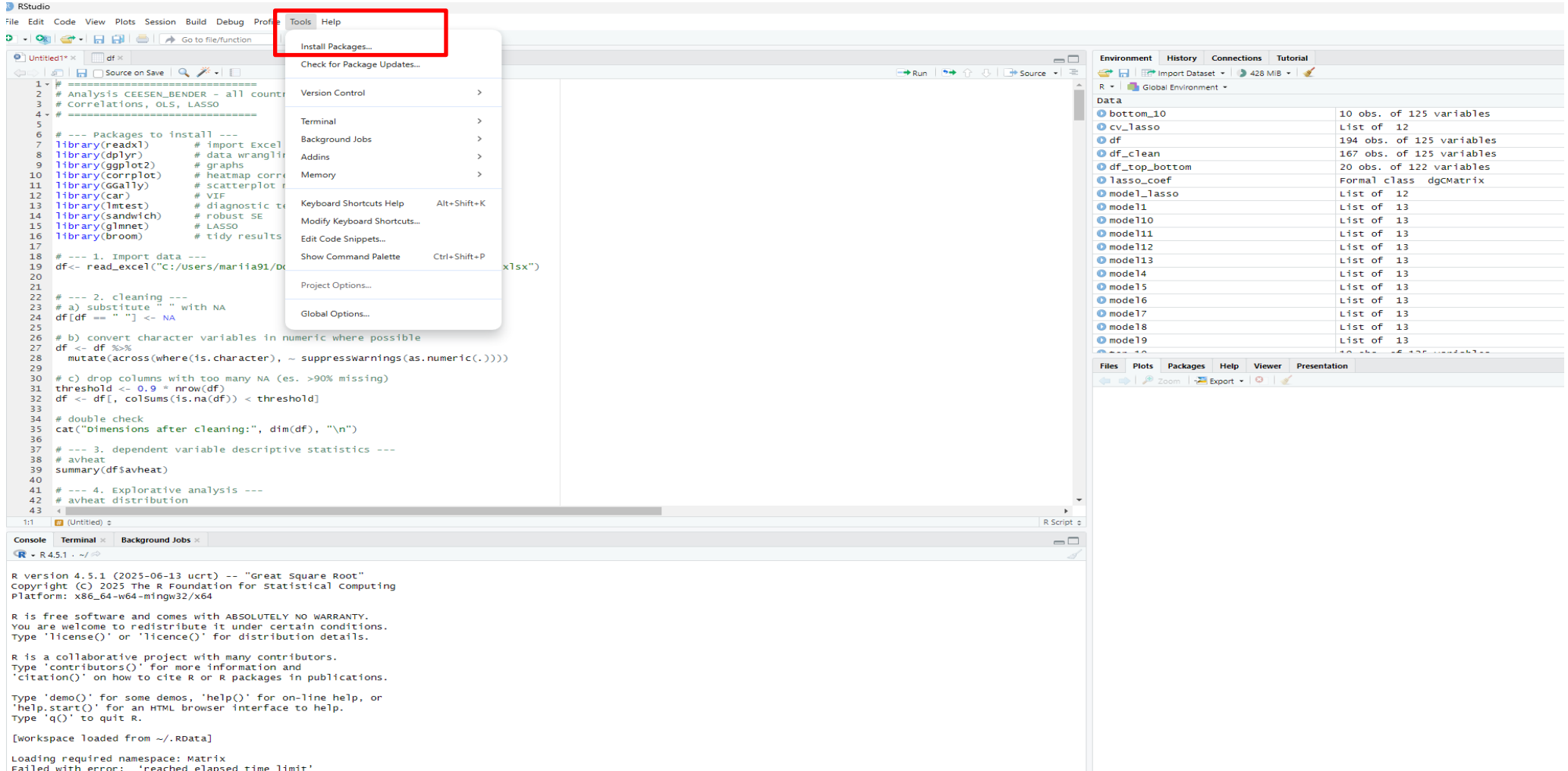
2. Select them all at once and run by either clicking “Run” command in the right hand part of the script window or as a combination of “Ctrl+Enter”



Installation of additional R packages for the analysis (2)

If these packages have not been previously installed, they should be installed one by one manually by clicking on

Tools – Install Packages

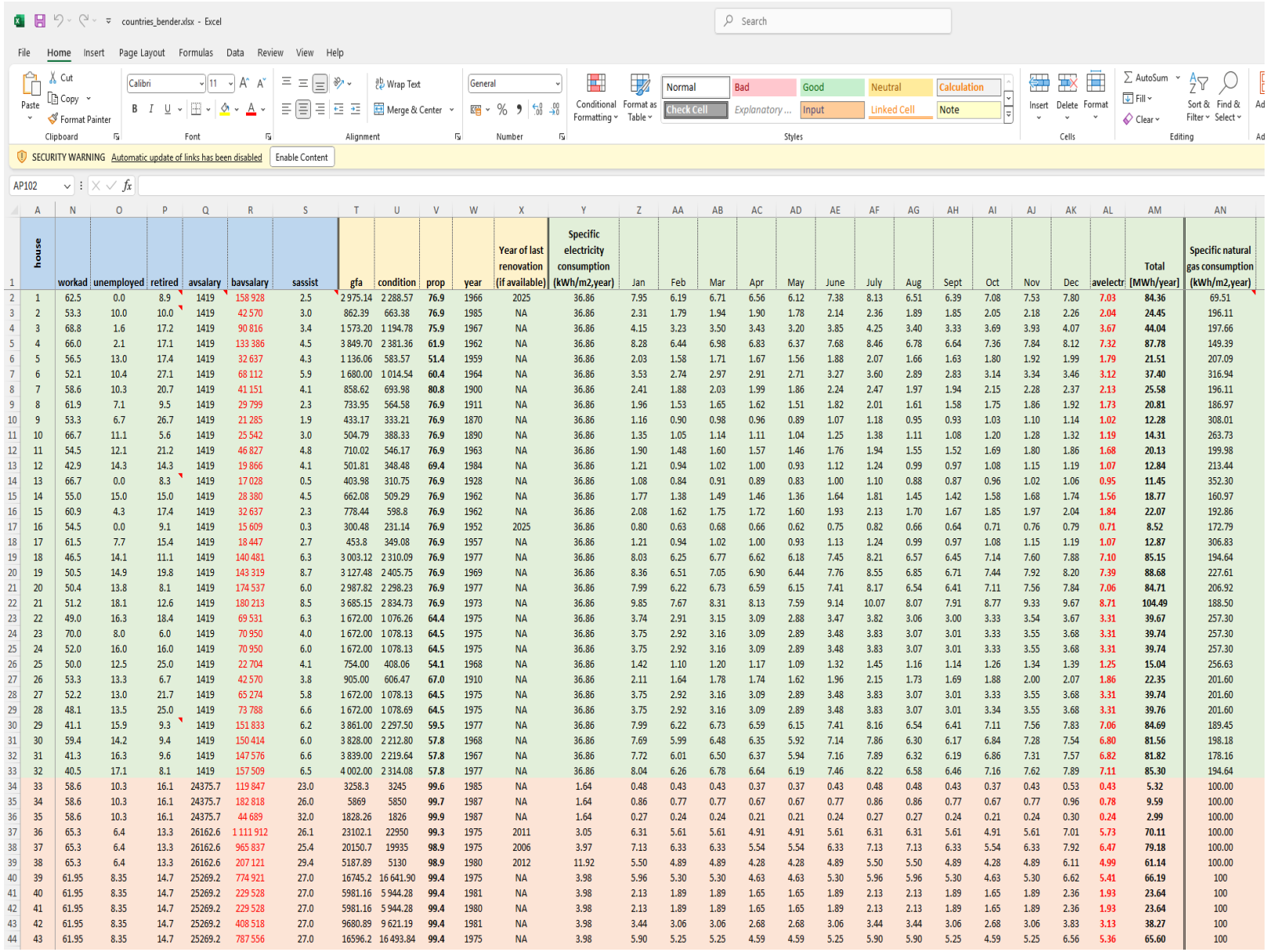


File preparation and storage

1. Save Excel file called "***countries_bender.xlsx***" with coded data* on *disk C* of your computer preferably under "**Documents**" as **.xlsx** type.

2. Make a separate folder. For example, called "**Projects**" under "**Documents**" and save Excel file into this folder.

* How the data has been coded is shown in the file called "Codebook"


countries_bender.xlsx - Excel

	house	workad	unemployed	retired	avsalary	bavsalary	sassist	gfa	condition	prop	year	Year of last renovation (if available)	Specific electricity consumption (kWh/m2,year)	Jan	Feb	Mar	Apr	May	June	July	Aug	Sept	Oct	Nov	Dec	avelectr	Total [MWh/year]	Specific natural gas consumption (kWh/m2,year)
2	1	62.5	0.0	8.9	1419	158 928	2.5	2 975.14	2 288.57	76.9	1966	2025	36.86	7.95	6.19	6.71	6.56	6.12	7.38	8.13	6.51	6.39	7.08	7.53	7.80	7.03	84.36	69.51
3	2	53.3	10.0	10.0	1419	42 570	3.0	862.39	663.38	76.9	1985	NA	36.86	2.31	1.79	1.94	1.90	1.78	2.14	2.36	1.89	1.85	2.05	2.18	2.26	2.04	24.45	196.11
4	3	68.8	1.6	17.2	1419	90 816	3.4	1 573.20	1 194.78	75.9	1967	NA	36.86	4.15	3.23	3.50	3.43	3.20	3.85	4.25	3.40	3.33	3.69	3.93	4.07	3.67	44.04	197.66
5	4	66.0	2.1	17.1	1419	133 386	4.5	3 849.70	2 381.36	61.9	1962	NA	36.86	8.28	6.44	6.98	6.83	6.37	7.68	8.46	6.78	6.64	7.36	7.84	8.12	7.32	87.78	149.39
6	5	56.5	13.0	17.4	1419	32 637	4.3	1 136.06	583.57	51.4	1959	NA	36.86	2.03	1.58	1.71	1.67	1.56	1.88	2.07	1.66	1.63	1.80	1.92	1.99	1.79	21.51	207.09
7	6	52.1	10.4	27.1	1419	68 112	5.9	1 680.00	1 014.54	60.4	1964	NA	36.86	3.53	2.74	2.97	2.91	2.71	3.27	3.60	2.89	2.83	3.14	3.34	3.46	3.12	37.40	316.94
8	7	58.6	10.3	20.7	1419	41 151	4.1	858.62	693.98	80.8	1900	NA	36.86	2.41	1.88	2.03	1.99	1.86	2.24	2.47	1.97	1.94	2.15	2.28	2.37	2.13	25.58	196.11
9	8	61.9	7.1	9.5	1419	29 799	2.3	733.95	564.58	76.9	1911	NA	36.86	1.96	1.53	1.65	1.62	1.51	1.82	2.01	1.61	1.58	1.75	1.86	1.92	1.73	20.81	186.97
10	9	53.3	6.7	26.7	1419	21 285	1.9	433.17	333.21	76.9	1870	NA	36.86	1.16	0.90	0.98	0.96	0.89	1.07	1.18	0.95	0.93	1.03	1.10	1.14	1.02	12.28	308.01
11	10	66.7	11.1	5.6	1419	25 542	3.0	504.79	388.33	76.9	1890	NA	36.86	1.35	1.05	1.14	1.11	1.04	1.25	1.38	1.11	1.08	1.20	1.28	1.32	1.19	14.31	263.73
12	11	54.5	12.1	21.2	1419	46 827	4.8	710.02	546.17	76.9	1963	NA	36.86	1.90	1.48	1.60	1.57	1.46	1.76	1.94	1.55	1.52	1.69	1.80	1.86	1.68	20.13	199.98
13	12	42.9	14.3	14.3	1419	19 866	4.1	501.81	348.48	69.4	1984	NA	36.86	1.21	0.94	1.02	1.00	0.93	1.12	1.24	0.99	0.97	1.08	1.15	1.19	1.07	12.84	213.44
14	13	66.7	0.0	8.3	1419	17 028	0.5	403.98	310.75	76.9	1928	NA	36.86	1.08	0.84	0.91	0.89	0.83	1.00	1.10	0.88	0.87	0.96	1.02	1.06	0.95	11.45	352.30
15	14	55.0	15.0	15.0	1419	28 380	4.5	662.08	509.29	76.9	1962	NA	36.86	1.77	1.38	1.49	1.46	1.36	1.64	1.81	1.45	1.42	1.58	1.68	1.74	1.56	18.77	160.97
16	15	60.9	4.3	17.4	1419	32 637	2.3	778.44	598.8	76.9	1962	NA	36.86	2.08	1.62	1.75	1.72	1.60	1.93	2.13	1.70	1.67	1.85	1.97	2.04	1.84	22.07	192.86
17	16	54.5	0.0	9.1	1419	15 609	0.3	300.48	231.14	76.9	1952	2025	36.86	0.80	0.63	0.68	0.66	0.62	0.75	0.82	0.66	0.64	0.71	0.76	0.79	0.71	8.52	172.79
18	17	61.5	7.7	15.4	1419	18 447	2.7	453.8	349.08	76.9	1957	NA	36.86	1.21	0.94	1.02	1.00	0.93	1.13	1.24	0.99	0.97	1.08	1.15	1.19	1.07	12.87	306.83
19	18	46.5	14.1	11.1	1419	140 481	6.3	3 003.12	2 310.09	76.9	1977	NA	36.86	8.03	6.25	6.77	6.62	6.18	7.45	8.21	6.57	6.45	7.14	7.60	7.88	7.10	85.15	194.64
20	19	50.5	14.9	19.8	1419	143 319	8.7	3 127.48	2 405.75	76.9	1969	NA	36.86	8.36	6.51	7.05	6.90	6.44	7.76	8.55	6.85	6.71	7.44	7.92	8.20	7.39	88.68	227.61
21	20	50.4	13.8	8.1	1419	174 537	6.0	2 987.82	2 298.23	76.9	1977	NA	36.86	7.99	6.22	6.73	6.59	6.15	7.41	8.17	6.54	6.41	7.11	7.56	7.84	7.06	84.71	206.92
22	21	51.2	18.1	12.6	1419	180 213	8.5	3 685.15	2 834.73	76.9	1973	NA	36.86	9.85	7.67	8.31	8.13	7.59	9.14	10.07	8.07	7.91	8.77	9.33	9.67	8.71	104.49	188.50
23	22	49.0	16.3	18.4	1419	69 531	6.3	1 672.00	1 076.26	64.4	1975	NA	36.86	3.74	2.91	3.15	3.09	2.88	3.47	3.82	3.06	3.00	3.33	3.54	3.67	3.31	39.67	257.30
24	23	70.0	8.0	6.0	1419	70 950	4.0	1 672.00	1 078.13	64.5	1975	NA	36.86	3.75	2.92	3.16	3.09	2.89	3.48	3.83	3.07	3.01	3.33	3.55	3.68	3.31	39.74	257.30
25	24	52.0	16.0	16.0	1419	70 950	6.0	1 672.00	1 078.13	64.5	1975	NA	36.86	3.75	2.92	3.16	3.09	2.89	3.48	3.83	3.07	3.01	3.33	3.55	3.68	3.31	39.74	257.30
26	25	50.0	12.5	25.0	1419	22 704	4.1	754.00	408.06	54.1	1968	NA	36.86	1.42	1.10	1.20	1.17	1.09	1.32	1.45	1.16	1.14	1.26	1.34	1.39	1.25	15.04	256.63
27	26	53.3	13.3	6.7	1419	42 570	3.8	905.00	606.47	67.0	1910	NA	36.86	2.11	1.64	1.78	1.74	1.62	1.96	2.15	1.73	1.69	1.88	2.00	2.07	1.86	22.35	201.60
28	27	52.2	13.0	21.7	1419	65 274	5.8	1 672.00	1 078.13	64.5	1975	NA	36.86	3.75	2.92	3.16	3.09	2.89	3.48	3.83	3.07	3.01	3.33	3.55	3.68	3.31	39.74	201.60
29	28	48.1	13.5	25.0	1419	73 788	6.6	1 672.00	1 078.69	64.5	1975	NA	36.86	3.75	2.92	3.16	3.09	2.89	3.48	3.83	3.07	3.01	3.34	3.55	3.68	3.31	39.76	201.60
30	29	41.1	15.9	9.3	1419	151 833	6.2	3 861.00	2 297.50	59.5	1977	NA	36.86	7.99	6.22	6.73	6.59	6.15	7.41	8.16	6.54	6.41	7.11	7.56	7.83	7.06	84.69	189.45
31	30	59.4	14.2	9.4	1419	150 414	6.0	3 828.00	2 212.80	57.8	1968	NA	36.86	7.69	5.99	6.48	6.35	5.92	7.14	7.86	6.30	6.17	6.84	7.28	7.54	6.80	81.56	198.18
32	31	41.3	16.3	9.6	1419	147 576	6.6	3 839.00	2 219.64	57.8	1967	NA	36.86	7.72	6.01	6.50	6.37	5.94	7.16	7.89	6.32	6.19	6.86	7.31	7.57	6.82	81.82	178.16
33	32	40.5	17.1	8.1	1419	157 509	6.5	4 002.00	2 314.08	57.8	1977	NA	36.86	8.04	6.26	6.78	6.64	6.19	7.46	8.22	6.58	6.46	7.16	7.62	7.89	7.11	85.30	194.64
34	33	58.6	10.3	16.1	24375.7	119 847	23.0	3258.3	3245	99.6	1985	NA	1.64	0.48	0.43	0.43	0.37	0.37	0.43	0.48	0.48	0.43	0.37	0.43	0.53	0.43	5.32	100.00
35	34	58.6	10.3	16.1	24375.7	182 818	26.0	5869	5850	99.7	1987	NA	1.64	0.86	0.77	0.77	0.67	0.67	0.77	0.86	0.86	0.77	0.67	0.77	0.96	0.78	9.59	100.00
36	35	58.6	10.3	16.1	24375.7	44 689	32.0	1828.26	1826	99.9	1987	NA	1.64	0.27	0.24	0.24	0.21	0.21	0.24	0.27	0.27	0.24	0.21	0.24	0.30	0.24	2.99	100.00
37	36	65.3	6.4	13.3	26162.6	1 111 912	26.1	23102.1	22950	99.3	1975	2011	3.05	6.31	5.61	5.61	4.91	4.91	5.61	6.31	6.31	5.61	4.91	5.61	7.01	5.73	70.11	100.00
38	37	65.3	6.4	13.3	26162.6	965 837	25.4	20150.7	19935	98.9	1975	2006	3.97	7.13	6.33	6.33	5.54	5.54	6.33	7.13	7.13	6.33	5.54	6.33	7.92	6.47	79.18	100.00
39	38	65.3	6.4	13.3	26162.6	207 121	29.4	5187.89	5130	98.9	1980	2012	11.92	5.50	4.89	4.89	4.28	4.28	4.89	5.50	5.50	4.89	4.28	4.89	6.11	4.99	61.14	100.00
40	39	61.95	8.35	14.7	25269.2	774 921	27.0	16745.2	16 641.90	99.4	1975	NA	3.98	5.96	5.30	5.30	4.63	4.63	5.30	5.96	5.96	5.30	4.63	5.30	6.62	5.41	66.19	100
41	40	61.95	8.35	14.7	25269.2	229 528	27.0	5981.16	5 944.28	99.4	1981	NA	3.98	2.13	1.89	1.89	1.65	1.65	1.89	2.13	2.13	1.89	1.65	1.89	2.36	1.93	23.64	100
42	41	61.95	8.35	14.7	25269.2	229 528	27.0	5981.16	5 944.28	99.4	1980	NA	3.98	2.13	1.89	1.89	1.65	1.65	1.89	2.13	2.13	1.89	1.65	1.89	2.36	1.93	23.64	100
43	42	61.95	8.35	14.7	25269.2	408 518	27.0	9680.89	9 621.19	99.4	1981	NA	3.98	3.44	3.06	3.06	2.68	2.68	3.06	3.44	3.44	3.06	2.68	3.06	3.83	3.13	38.27	100
44	43	61.95	8.35	14.7	25269.2	787 556	27.0	16596.2	16 493.84	99.4	1975	NA	3.98	5.90	5.25	5.25	4.59	4.59	5.25	5.90	5.90	5.25	4.59	5.25	6.56	5.36	65.60	100

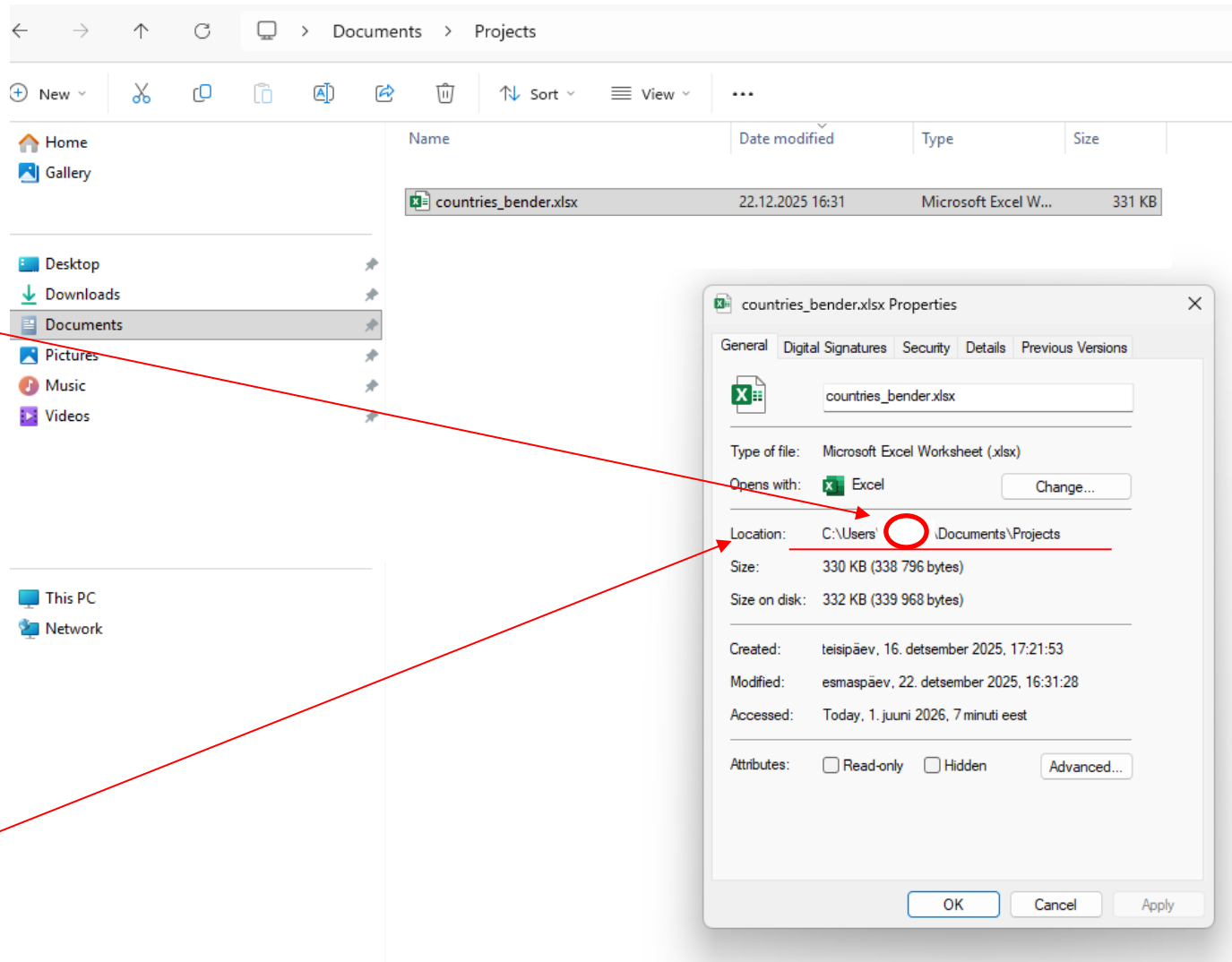
Loading dataset into R environment (1)

1. Copy all the codes from the file called "***R codes.txt***" into script window in R programme.
2. Execute the command (Ctrl+Enter):

df<- read_excel("C:/Users/YOUR NAME/Documents/Projects/countries_bender.xlsx")

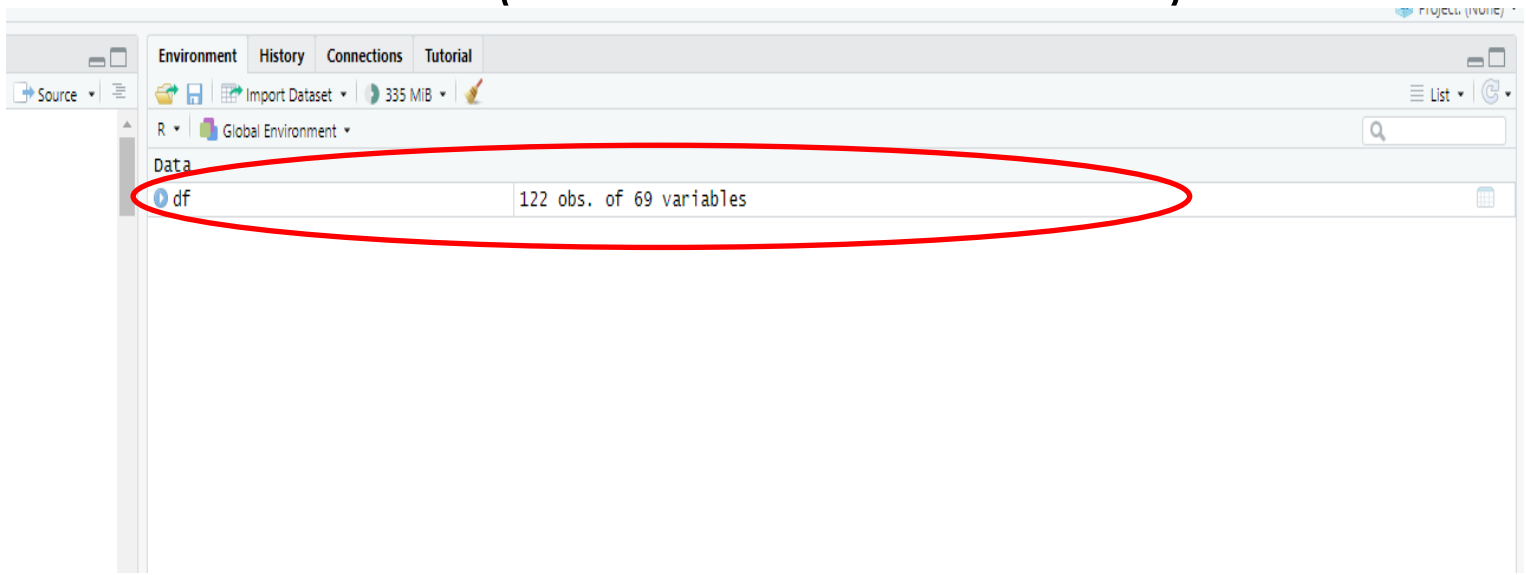
NB! The path to your file in " " will vary, you have to check it under "***countries_bender.xlsx***" file's "Properties"

For that: Go to Documents → folder "Projects" → right click on the file "*countries_bender.xlsx*" → Properties → under "General", you will see the correct location of your file.



Loading dataset into R environment (2)

3. If there were no errors running these commands (due to missing parenthesis for example), then a dataset named "df" should appear in the top right corner window (environment window):



If you click on the dataset row, the dataset or its information is shown. For more details on how to use R programme, one can use a user-friendly online guidebook available here or manuals allocated on R website here.

4. Run all codes from "***R codes.txt***" file

CENTRAL & EASTERN EUROPEAN SUSTAINABLE ENERGY NETWORK

CEESEN-BENDER

Description of the model and results

Co-funded by the European Union

Dataset overview

Cross-sectional annual data has been collected by BENDER project partners in four countries – Croatia, Romania, Slovenia and Poland – to pilot the tool developed initially on Estonian data. Three groups of data are used:

Socio-economic data of MAB* residents	Technical characteristics of buildings	Energy data
• Number of (mostly) empty flats with 0 residents • Number of flats with 1 resident • Number of flats with 2 residents • Number of flats with 3 residents or more • Average living area per flat (m2) • Number of residents living in the MAB • Average living area per person (m2) • Percentage of owners living in the MAB (%) • Percentage of tenants living in the MAB (%) • Children below 18 of age (%) • Working adults 18-60 of age (%) • Unemployed adults 18-60 of age (%) • Pensioners above 60 of age (%) • Average salary per dwelling (€ gross) • Dwellings with residents receiving social assistance (%)	• GFA - Gross Floor Area (m2) • Conditioned area (m2) • Conditioned area % from GFA • Year of construction • Year of last renovation (if available)	• MAB electricity consumption [MWh] • MAB natural gas consumption [MWh] • MAB heating costs [€]

*MAB – multi-apartment buildings

Sample and Variables

The sample consists of 122 observations from four countries.

Independent variables:

- Country
- Household composition (number of dwellings with 0, 1, 2, 3 or more residents)
- Average living area per flat
- Average living area per person
- Ownership structure:
  - Share of owners (%)
  - Share of renters (%)
- Employment structure:
  - Children below 18 of age (%)
  - Working adults 18-60 of age (%)
  - Unemployed adults 18-60 of age (%)
  - Pensioners above 60 of age (%)
- Income (measured as average monthly salary per building (€ gross))
- Flats with residents receiving social assistance (%)
- GFA - Gross Floor Area (m2)
- Conditioned area (m2)
- Year of construction
- Average electricity consumption of building per year [MWh]
- Average natural gas consumption of building per year [MWh]

Dependent variable:

Average annual heating costs, €

Correlation analysis, OLS and LASSO

- *Correlation analysis* showed which independent variables correlate with each other → only one should be included

- 16 *Multiple Linear Regression (MLR) models* were tested to identify which factors are consistently significant across all the models

- *LASSO model* was applied to double check the results of OLS. When many predictors are involved LASSO identifies the most important from them.

Multicollinearity and Heteroscedasticity Tests

- VIF (Variance Inflation Factor) is applied to detect a multicollinearity problem in tested models. VIF between 1 and 5 is an acceptable range.

- Residuals
  - Check for linearity and homoscedasticity
  - Check for normality of residuals

- Robust Standard Error (test for heteroscedasticity)

LASSO model for Croatia, Poland and Romania

Significant factors	LASSO coefficient	Weights*
Average monthly salary per building	0.000002101748	0.05
Country	0.5015637	0.4
Percent of unemployed adults age 18-60	0.02386472	0.15
Number of (mostly) empty dwellings with 0 residents	0.06683785	0.3
Year of construction	0.006892765	0.1

*Weights are selected based on the magnitude of LASSO coefficients

Calculation of a weighted score:

country * weights[0.4] + resident0_std * weights[0.3] +
unemployed_std * weights[0.15] + year_std * weights[0.1] +
bavsalary_std * weights[0.05]

LASSO model for Slovenia

For Slovenia, factor "unemployed" has to be excluded from the analysis, since no observations for this factor are available.

Significant factors	LASSO coefficient	Weights*
Average monthly salary per building	0.000002101748	0.1
Country	0.5015637	0.4
Number of (mostly) empty dwellings with 0 residents	0.06683785	0.3
Year of construction	0.006892765	0.2

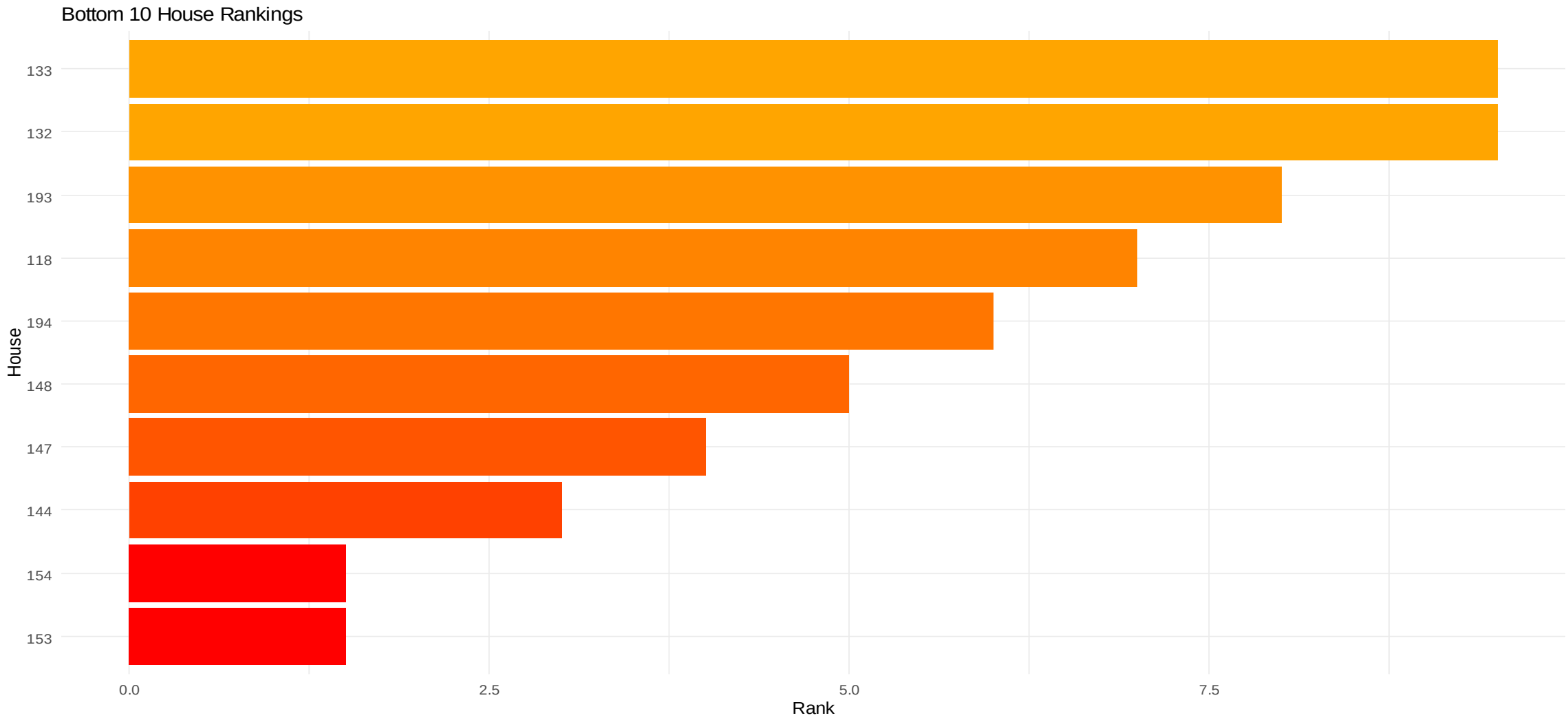
*Weights are selected based on the magnitude of LASSO coefficients

Calculation of a weighted score:

country_std * weights[0.4] + resident0_std * weights[0.3] +
year_std * weights[0.2] + bavsalary_std *weights[0.1]

Houses which require renovation first – lowest rank score

The outcome of the model is ranking of the buildings as it is exampled below. Where the houses that collected less points (dark red) require renovation first.



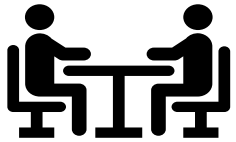
Limitations of the model – room for improvement!

- Small dataset!
  - Longitudinal data would be preferrable (e.g. 5 years)
  - More variables to include (e.g. some other technical characteristics of the buildings are omitted)
- Seasonality is not captured (cross-sectional data)
- Gas data was not available for all the buildings.
- Given a richer dataset, a different statistical model can be applied.
- Slovenian sample was not full.

Who can use the model?

Data owners!

Local authorities (e.g., for renovation of municipal buildings)



Financial intermediaries (e.g., KredEx in Estonia) for deciding on the grants

Energy agencies for data driven advise


CEE
SEN