AUTOMATIC DATA VALIDATION FOR SILC

Margherita Zuppardo

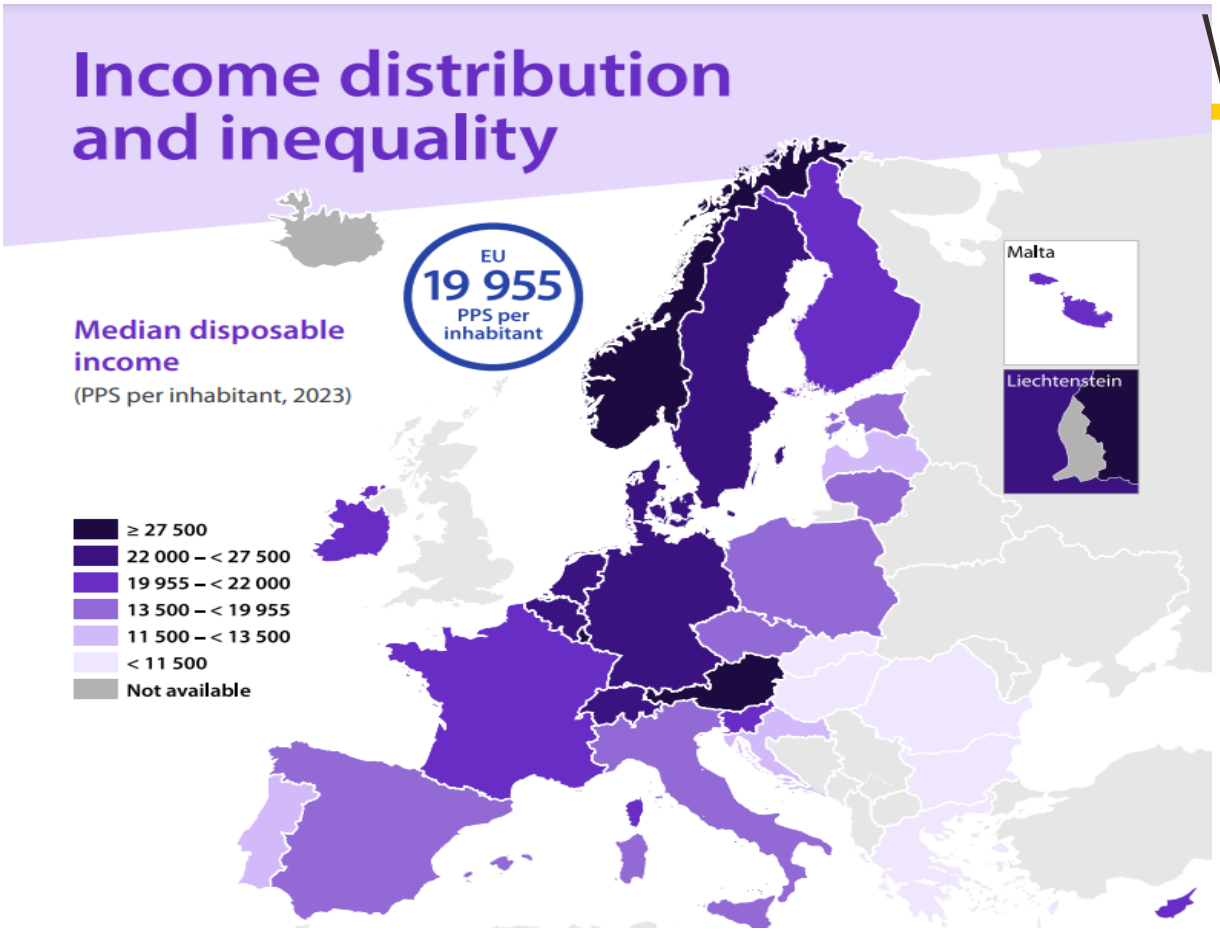
margherita.zuppardo@hagstofa.is

uRos 2024 ATHENS

urosconf.org

Statistics Iceland

What is SILC?



The EU statistics on income and living conditions (EU-SILC) aim to collect timely and comparable cross-sectional and longitudinal data on **income, poverty, social exclusion, and living conditions**.

EU-SILC is a household and individual data collection which output is harmonised as it is regulated by legislations. Around 90% of the data collection is made up of annual variables. The rest are either

[1,2] ec.europa.eu

The SILC Team in Iceland

Data collection

Data processing and analysis

The “doc65”

EUROPEAN COMMISSION
EUROSTAT

Directorate F: Social Statistics
Unit F-4: Quality of life

DocSILC065 (2023 operation)

METHODOLOGICAL GUIDELINES AND DESCRIPTION OF EU-SILC TARGET VARIABLES

2023 operation (Version 5)

SILC Variables

TABLE OF CONTENTS

Data processing in R

```r
year1='21' #year I am working on


#make the D-file
source('D_file_light.r')

#make the R-file
source('R_file.r')

#make the P-file
source('P_file_tax.r')

##make the H-file
source('H_file_no_tax.r')

source(paste('ad_hoc',year1,'.r',sep=''))

source(paste('rolling_modules_20',year1,'.r',sep=''))

source('make_reconciled2.r')

source('weights_integrated_bergros22.r')
if(year1 != year1_start) error= erro

source('Reorder_files3.r')
```

Data processing in R

```r
year1='21' #year I am working on


#make the D-file
source('D_file_light.r')

#make the R-file
source('R_file.r')

#make the P-file
source('P_file_tax.r')

##make the H-file
source('H_file_no_tax.r')

source(paste('ad_hoc',year1,'.r',sep=''))

source(paste('rolling_modules_20',year1,'.r',sep=''))

source('make_reconciled2.r')

source('weights_integrated_bergros22.r')
if(year1 != year1_start) error= erro

source('Reorder_files3.r')

##########################Check codes

#General checks on the variables
source('check_values.r')

#Aggragate consistency among years
source('check_values_weights.r')


#Total populations stemming from the weights
source('check_weights.r')


#Logical consistency checks (between varables and between years)
source('logical_checks.r')
```

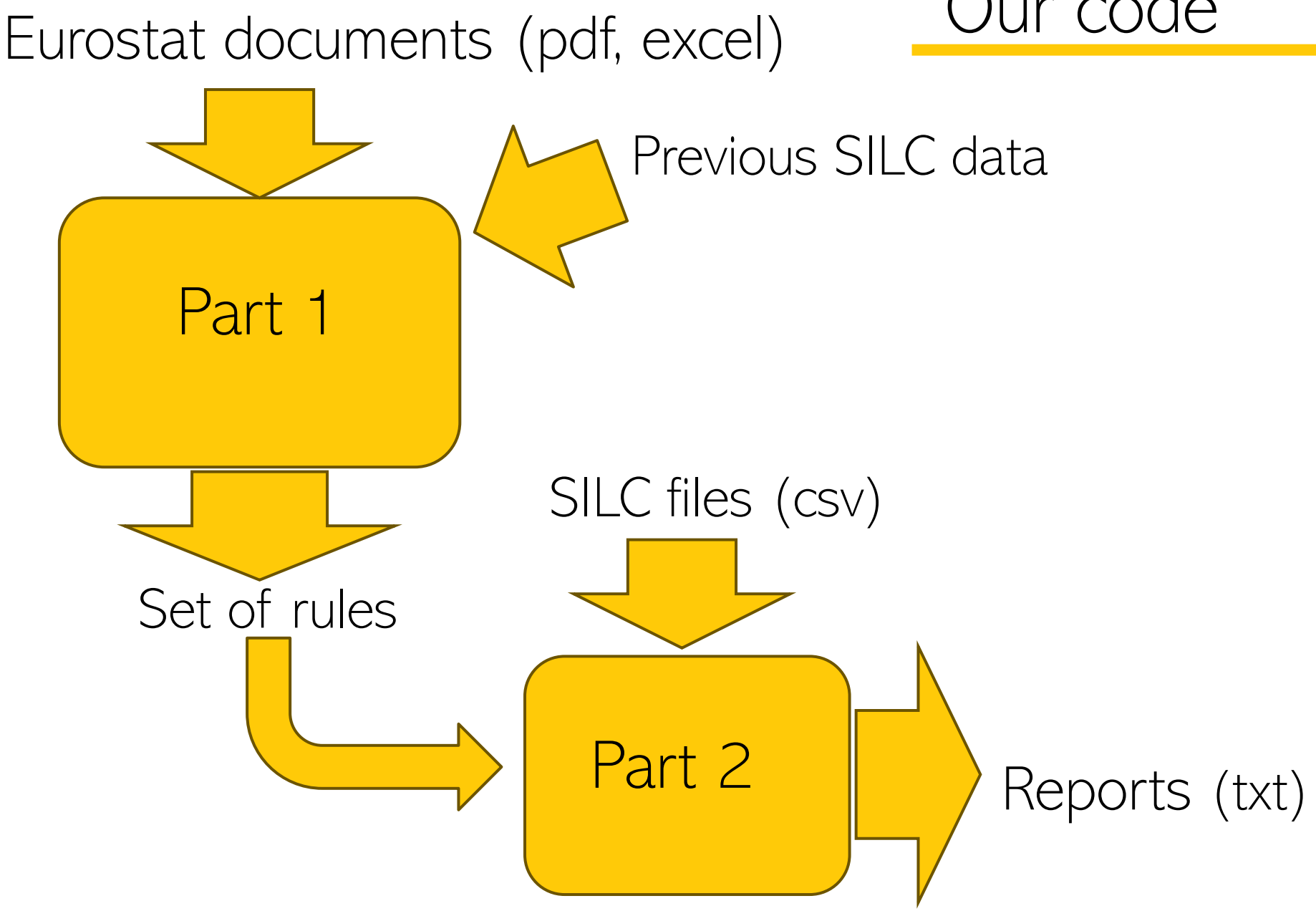
Data validation

```
##########################check codes

#General checks on the variables
source('check_values.r')
## missing columns
## too many missing values (above 20%)
## accepted values
## correct flags

#Aggragate consistency among years
source('check_values_weights.r')
## mean, median and sd of weighted variables consistent with previous year

#Total populations stemming from the weights
source('check_weights.r')


#Logical consistency checks (between varables and between years)
source('logical_checks.r')
```

Data validation

```
##########################check codes

#General checks on the variables
source('check_values.r')
## missing columns
## too many missing values (above 20%)
## accepted values
## correct flags

#Aggragate consistency among years
source('check_values_weights.r')
## mean, median and sd of weighted variables consistent with previous year

#Total populations stemming from the weights
source('check_weights.r')


#Logical consistency checks (between varables and between years)
source('logical_checks.r')
```

Our code


Eurostat documents (pdf, excel)
Previous SILC data
Part 1
SILC files (csv)
Set of rules
Part 2
Reports (txt)

Why we need automatic rule extraction

"Rules" change between years

RB090: Sex

Domain/Area	Basic data/Demograp
Transmission type	Early and regular
Reference period	Constant
Unit	All current household members (of any age) and former household members
Mode of collection	Household responden
Values	*1* Male *2* Female
Flags	*1* Filled *-1* Missing

```
(R_file1 |> filter(RB090_F %!in% c(1,-1)) |> nrow
```

2021

Personal Register (R-file)

RB090: SEX

Topic and detailed topic: Person and household characteristics / Demography
Variable type: Core Variable/First wave/ Annual
Unit: All current household members (of any age)
Reference period: Current
Mode of collection: Household respondent or registers
In use (period): Yes, since the first year of EU-SILC data collection
Series' differences: No changes

VALUES AND FORMAT

1	Male
2	Female

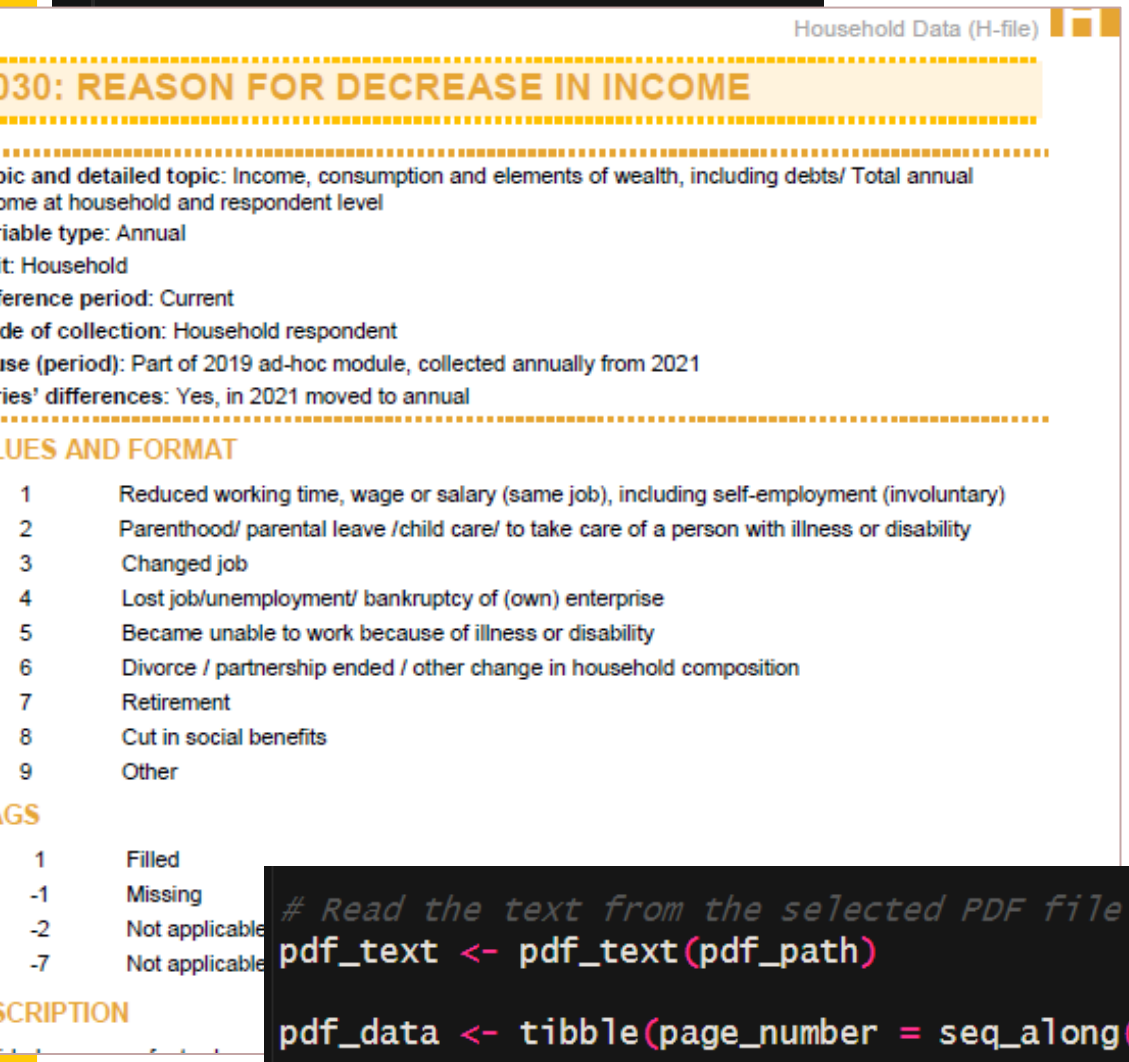
FLAGS

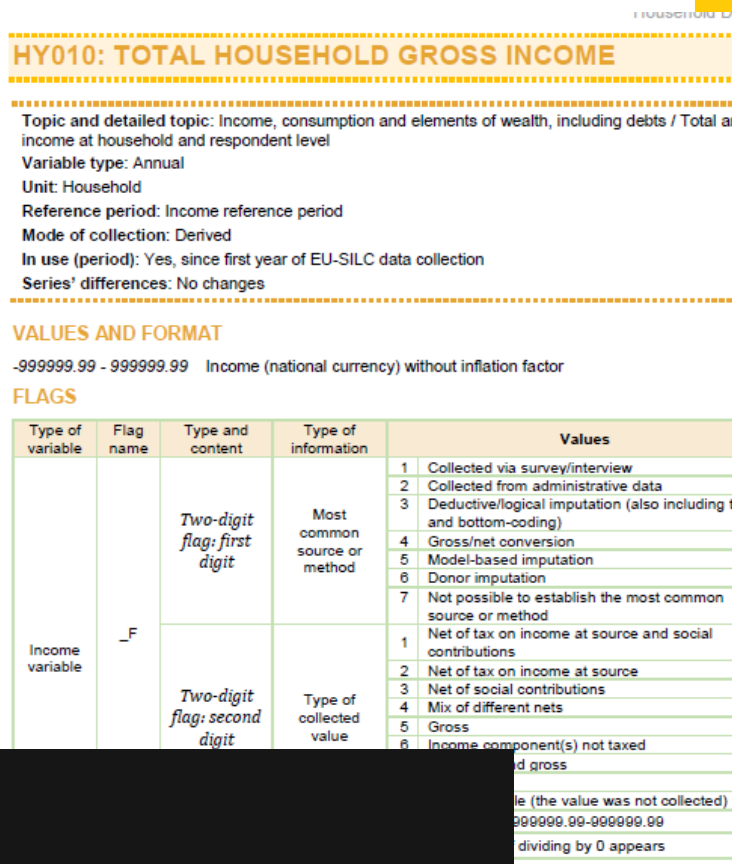
1	Main source is survey or interview
2	Main source is administrative data
3	Imputed
4	It is not possible to establish a main source
-1	Missing

```
(R_file1 |> filter(RB090_F %!in% c(1,2,3,4,-1)) |> nrow()) >0
```

```
library(pdftools)
```

Step 1: pdf text






```
# Read the text from the selected PDF file
pdf_text <- pdf_text(pdf_path)

pdf_data <- tibble(page_number = seq_along(pdf_text), text = pdf_text)
```

[3] https://cran.r-project.org/web/packages/pdftools

Sections

```r
#### now add the 'topic' column which says what the line is describing
mutate(
  # Identify lines that contain specific keywords
  is_values = str_detect(lines, paste0("^\\s*", keywords$Values_string, "\\s*$")),
  is_flags = str_detect(lines, paste0("^\\s*", keywords$Flags_string, "\\s*$")),
  is_description = str_detect(lines, paste0("^\\s*", keywords$Description_string, "\\s*$")),

  # Generate section ids based on occurrences of these keywords
  section_id = cumsum(is_values | is_flags | is_description)
) %>% ungroup() %>%
# Assign topic based on the section type detected
mutate(
  topic = case_when(
    is_values ~ "values",
    is_flags ~ "flags",
    is_description ~ "description",  # Optionally h
    TRUE ~ NA_character_
  )
) %>%
```


HI030: REASON FOR DECREASE IN INCOME

Topic and detailed topic: Income, consumption and elements of wealth, including debts/ Total annual income at household and respondent level
Variable type: Annual
Unit: Household
Reference period: Current
Mode of collection: Household respondent
In use (period): Part of 2019 ad-hoc module, collected annually from 2021
Series' differences: Yes, in 2021 moved to annual

VALUES AND FORMAT

1	Reduced working time, wage or salary (same job), including self-employment (involuntary)
2	Parenthood/ parental leave /child care/ to take care of a person with illness or disability
3	Changed job
4	Lost job/unemployment/ bankruptcy of (own) enterprise
5	Became unable to work because of illness or disability
6	Divorce / partnership ended / other change in household composition
7	Retirement
8	Cut in social benefits
9	Other

FLAGS

1	Filled
-1	Missing
-2	Not applicable (HI010 not equal to 3)
-7	Not applicable (HB010 not equal to 2021)

DESCRIPTION

HY010: TOTAL HOUSEHOLD GROSS INCOME

Topic and detailed topic: Income, consumption and elements of wealth, including debts / Total annual income at household and respondent level
Variable type: Annual
Unit: Household
Reference period: Income reference period
Mode of collection: Derived
In use (period): Yes, since first year of EU-SILC data collection
Series' differences: No changes

VALUES AND FORMAT

999999.99 – 999999.99 Income (national currency) without inflation factor

FLAGS

Type of variable	Flag name	Type and content	Type of information	Values	
Income variable	_F	*Two-digit flag: first digit*	Most common source or method	1	Collected via survey/interview
				2	Collected from administrative data
				3	Deductive/logical imputation (also including top- and bottom-coding)
				4	Gross/net conversion
				5	Model-based imputation
				6	Donor imputation
				7	Not possible to establish the most common source or method
		Two-digit flag: second digit	Type of collected value	1	Net of tax on income at source and social contributions
				2	Net of tax on income at source
				3	Net of social contributions
				4	Mix of different nets
				5	Gross
				6	Income component(s) not taxed
				7	Mix of net and gross
				8	Unknown
				9	Not applicable (the value was not collected)
	_IF	Imputation factor = collected value / recorded value *100		-999999.99-999999.99	
				1	If problem of dividing by 0 appears

Regular expressions

```
range_pattern <-
"(-?\\d+(?:\\.\\d+)?)\\s*(?:[-–—]|to)\\s*(-?\\d+(?:\\.\\d+)?)
```

```
(-?\\d+(?:\\.\\d+)?)
```

A negative integer, or a positive integer, or a real number

```
\\s*(?:[-–—]|to)\\s*
```

A hyphen (-), en-dash (–), or em-dash (—), the word „to“

VALUES AND FORMAT

0–99 Number of hours of education during a typical week

VALUES AND FORMAT

1 - 999999.99 Benefits
- 999999.99 - -1 Losses

VALUES AND FORMAT

1 – 31 Day

VALUES AND FORMAT

BE Belgium
BG Bulgaria
CZ Czechia
DK Denmark
DE Germany

Format: Four-digit number, no decimals

VALUES AND FORMAT

NUTS 2 digits

VALUES AND FORMAT

Household from previous wave
1 At the same address
2 Entire household mo
Household no longer in-scope
3 Entire household mo
4 Household moved o

Extracting the rule[4]

```r
library(stringr)
```

```r
var = 'HY010'
text <- pdf_data |> filter(variable_name == var  & topic == 'values' & lines != keywords$Values_strin
  pull(lines) |> paste0()
```

```r
# Detect ranges like "0-9", or "-9999.999 - 99999.99", or "1.5 to 3.5"
ranges <- stringr::str_match_all(text, range_pattern)
print(ranges)
```

```
> print(ranges)
[[1]]
     [,1]                        [,2]          [,3]
[1,] "-999999.99 - 999999.99" "-999999.99" "999999.99"
```

```r
# Transforms numbers to logical conditions
if (length(ranges[[1]]) > 0) {
  condition <-  paste('value >= ',  ranges[[1]][, 2], " & value <= ", ranges[[1]][, 3])
}
print(condition)
```

Result

```
> print(condition)
[1] "value >=  -999999.99  & value <=  999999.99"
```

HY010: TOTAL HOUSEHOLD GROSS INCOME

elements of wealth, including d

Reference period: Income reference period

Mode of collection: Derived

In use (period): Yes, since first year of EU-SILC data collection

Series' differences: No changes

VALUES AND FORMAT

-999999.99 - 999999.99 Income (national currency) without inflation factor

FLAGS

	Values
1	Collected via survey/interview
2	Collected from administrative
3	Deductive/logical imputation (a and bottom-coding)
4	Gross/net conversion
5	Model-based imputation
6	Donor imputation
7	Not possible to establish the m source or method

flag: first digit — source or method

1	Net of tax on income at source contributions
2	Net of tax on income at source
3	Net of social contributions
4	Mix of different nets
5	Gross
6	Income component(s) not taxe

flag: second digit — type of collected value

Other validation

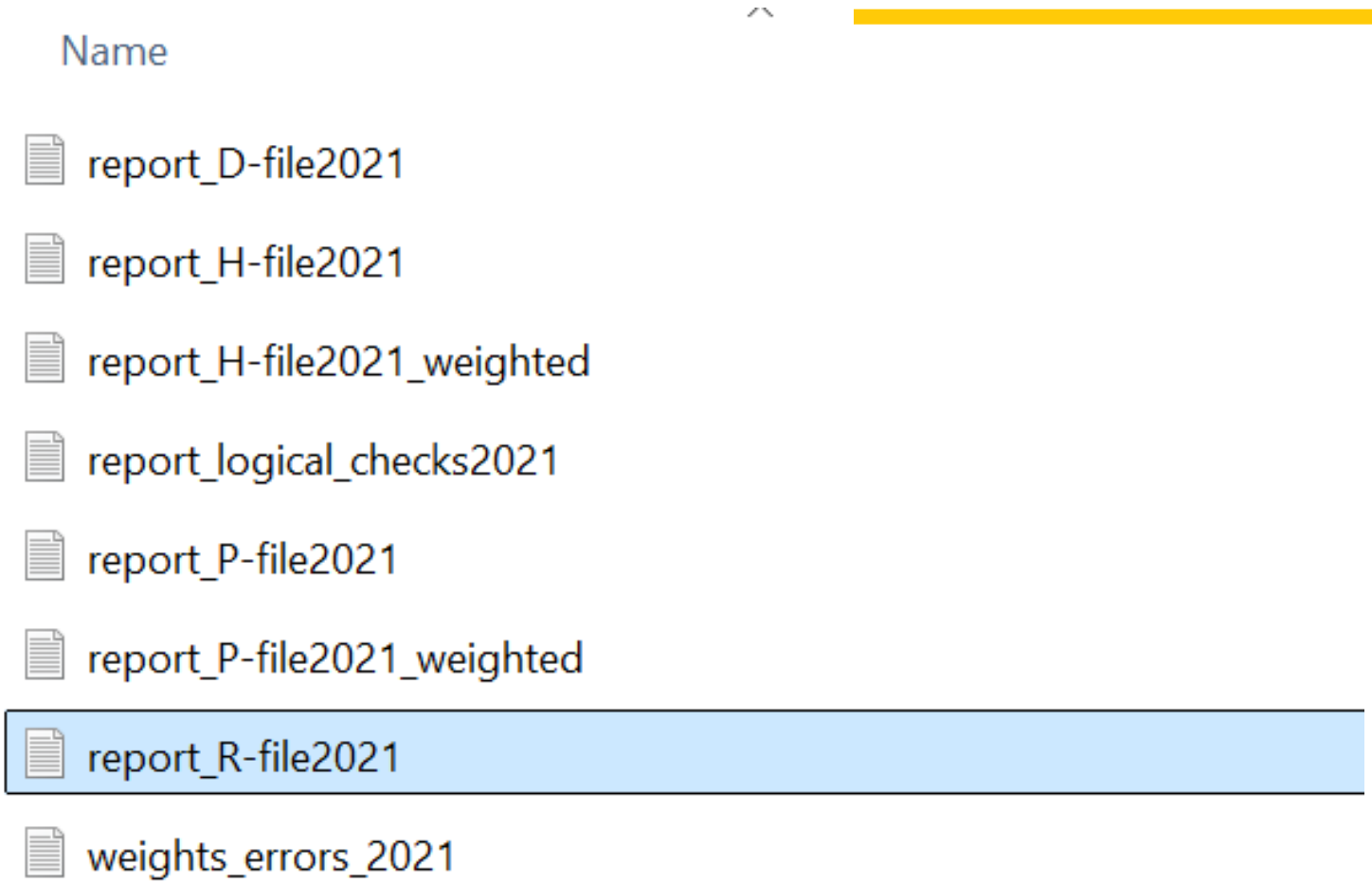
```
##############################check codes

#General checks on the variables
source('check_values.r')
## missing columns
## too many missing values (above 20%)
## accepted values
## correct flags

#Aggragate consistency among years
source('check_values_weights.r')
## mean, median and sd of weighted variables consistent with previou

#Total populations stemming from the weights
source('check_weights.r')


#Logical consistency checks (between varables and between years)
source('logical_checks.r')
```

https://cran.r-project.org/web/packages/rvest

https://cran.r-project.org/web/packages/pxweb

https://cran.r-project.org/web/packages/readxl

Final reports



Final reports

report_R-file2021 - Notepad

File Edit View

```
There are 4 variables with errors or warnings in  R-file

There are 2 variables with too many missing values in R-file

There are 1 variables with values that are not allowed in R-file

There are 1 variables with wrong flags in R-file

************************************ RK020****************************************
                            AGE OF CHILD WHO IS NOT HOUSEHOLD MEMBER

Warning! Variable: RK020. 1 values outside the desired range. Example: 98

************************************ RK070****************************************
                            NUMBER OF NIGHTS PER MONTH THE CHILD SPENDS IN THE HOUS

Warning! Variable  RK070  has  22 % missing values.

************************************ RK080****************************************
                            LEGAL CHILD CUSTODY SITUATION

Warning! Variable  RK080  has  22 % missing values.

************************************ RB245****************************************
                            RESPONDENT STATUS

Years valid:  >= 2006 .  Error: Flag values NA not allowed. Expected value: 1.
```

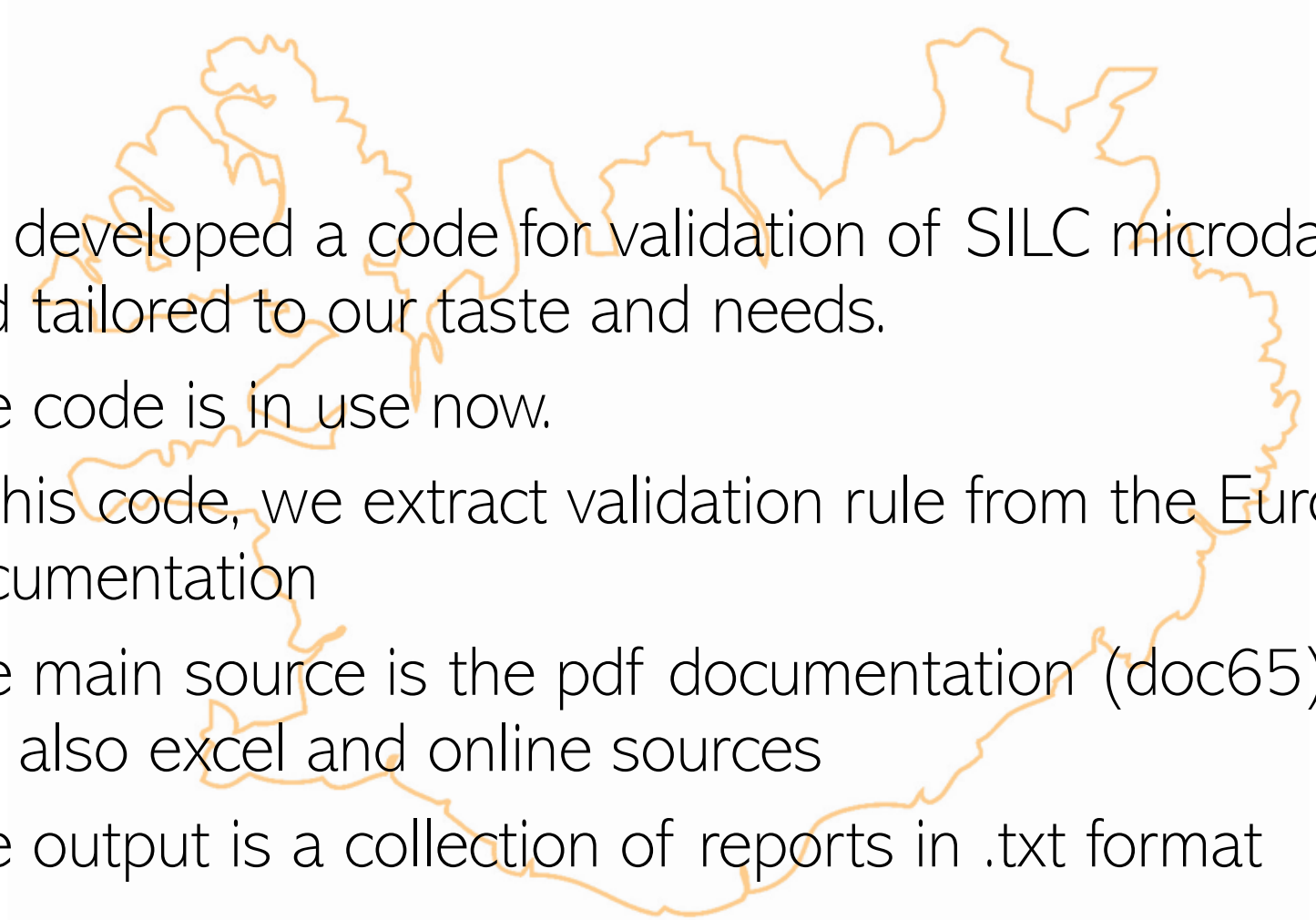
Final reports

report_logical_checks2021 - Notepad

File Edit View

```
------------------- as.numeric(PY010G )<= 0 & (PL211A == 1 | PL211A == 2) --------------------

------------------- Person is an employee (PL211A = 1 or 2) but has no income from employment (PY010

29 row(s) affected

"Temp_id2"         "PY010G" "PY010G_F"        "PL211A" "PL211A_F"
"03471700012020"   "0"      "0"      "2"      "1"
"05341700012020"   "0"      "0"      "1"      "1"
"08451700012020"   "0"      "0"      "2"      "1"
"09331700012020"   "0"      "0"      "1"      "1"
"02811800012020"   "0"      "0"      "1"      "1"
"04311700012020"   "0"      "0"      "1"      "1"
"12321800012020"   "0"      "0"      "1"      "1"
"02181800012020"   "0"      "0"      "1"      "1"
"03041900012020"   "0"      "0"      "1"      "1"
"05041900012020"   "0"      "0"      "2"      "1"



------------------- as.integer(PL060) +as.integer( PL100) >60 --------------------

------------------- Nr of hours usually worked is more than 60 hours per week --------------------

64 row(s) affected
```

Conclusions



- We developed a code for validation of SILC microdata and tailored to our taste and needs.
- The code is in use now.
- In this code, we extract validation rule from the Eurostat documentation
- The main source is the pdf documentation (doc65), but also excel and online sources
- The output is a collection of reports in .txt format

Conclusions

Pros:

- It works (most of the time)
- R-based pipeline simplifies SILC validation.
- The validation works on different years, with little maintenance.
- It takes minutes to run
- Automation reduces human error
- All packages (and R) are free

Cons:

- Other programming languages (Python) may offer more tools
- It took a while to develop the code initially.
- It depends on Eurostat documentation uniformity, requires robust design.
- It still needs some manual coding (logical checks)
- It is dependent on external packages
- Not as complete as Eurostat SAS code

THANK YOU!

[1] ec.europa.eu/eurostat/web/products-key-figures/w/ks-01-24-001

[2] ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions

[3] https://cran.r-project.org/web/packages/pdftools

[4] https://cran.r-project.org/web/packages/stringr/

Code coming soon

https://github.com/MargheritaZ/SilcValidation

Statistics Iceland

margherita.zuppardo@hagstofa.is

```
range_pattern <- "(\\d+(?:\\.\\d+)?|\\-\\d+(?:\\.\\d+)?)\\s*(?:[-–—]|to)\\s*(\\d+(?:\\.\\d+)?|\\-\\d+(?:\\.\\d+)?)"

discrete_pattern <- "(?<=^|\\n)\\s*(\\d+)\\.?\\s+(?!-)(?!\\.).*"

open_range_pattern <- "(\\d+)\\+"
```