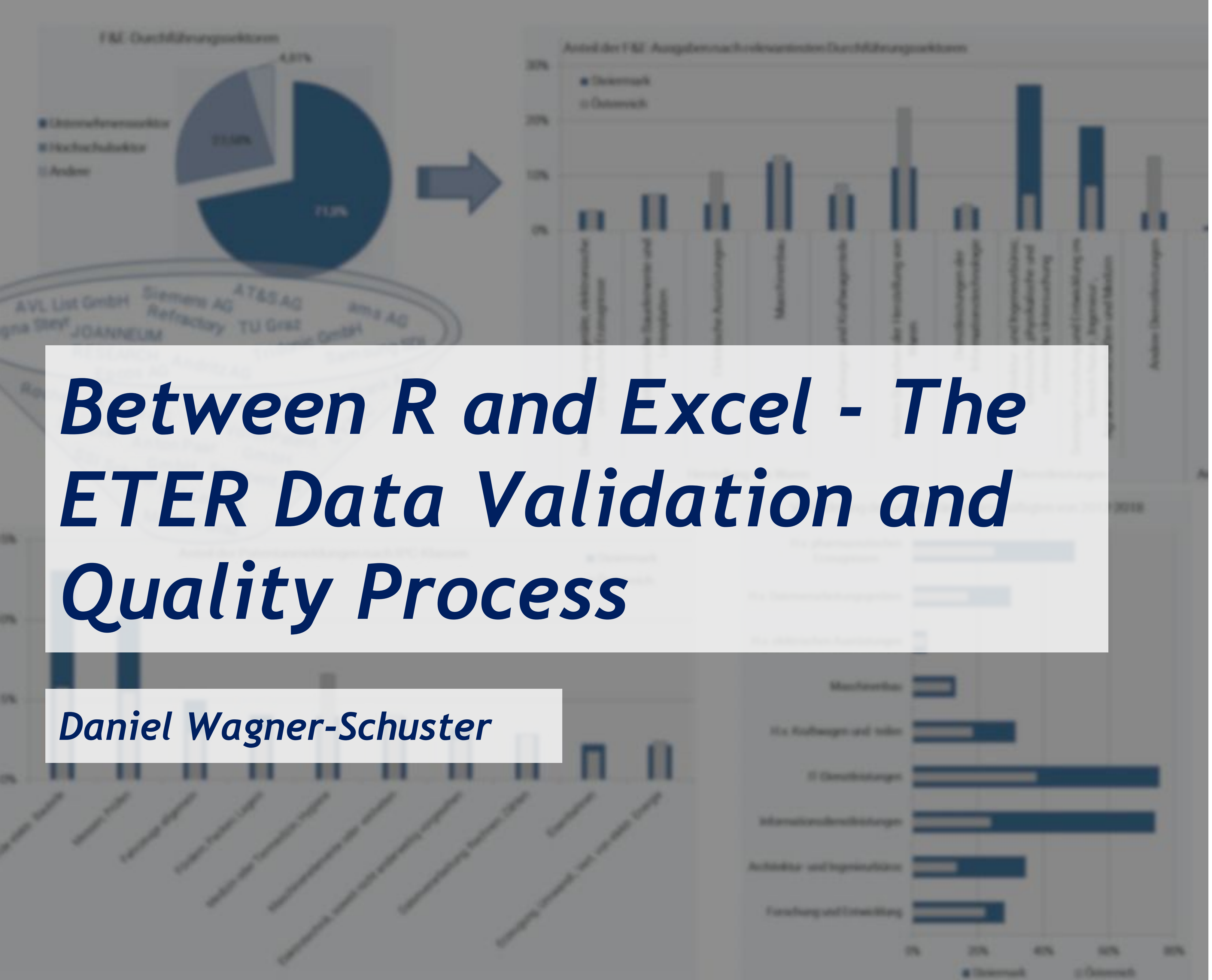


Between R and Excel - The ETER Data Validation and Quality Process

Daniel Wagner-Schuster

JOANNEUM RESEARCH
POLICIES

European Tertiary Education Register

ETER
EUROPEAN TERTIARY EDUCATION REGISTER

- Collects and provides data on higher education institutions in Europe
  - at the institutional level
  - 2013-2019
    - Tender for another 3 years
  - 37 countries
    - ~ 3,000 institutions per year
  - currently data for 6 academic years
    - 17,285 observations for 618 variables
  - eter-project.com

Some challenges

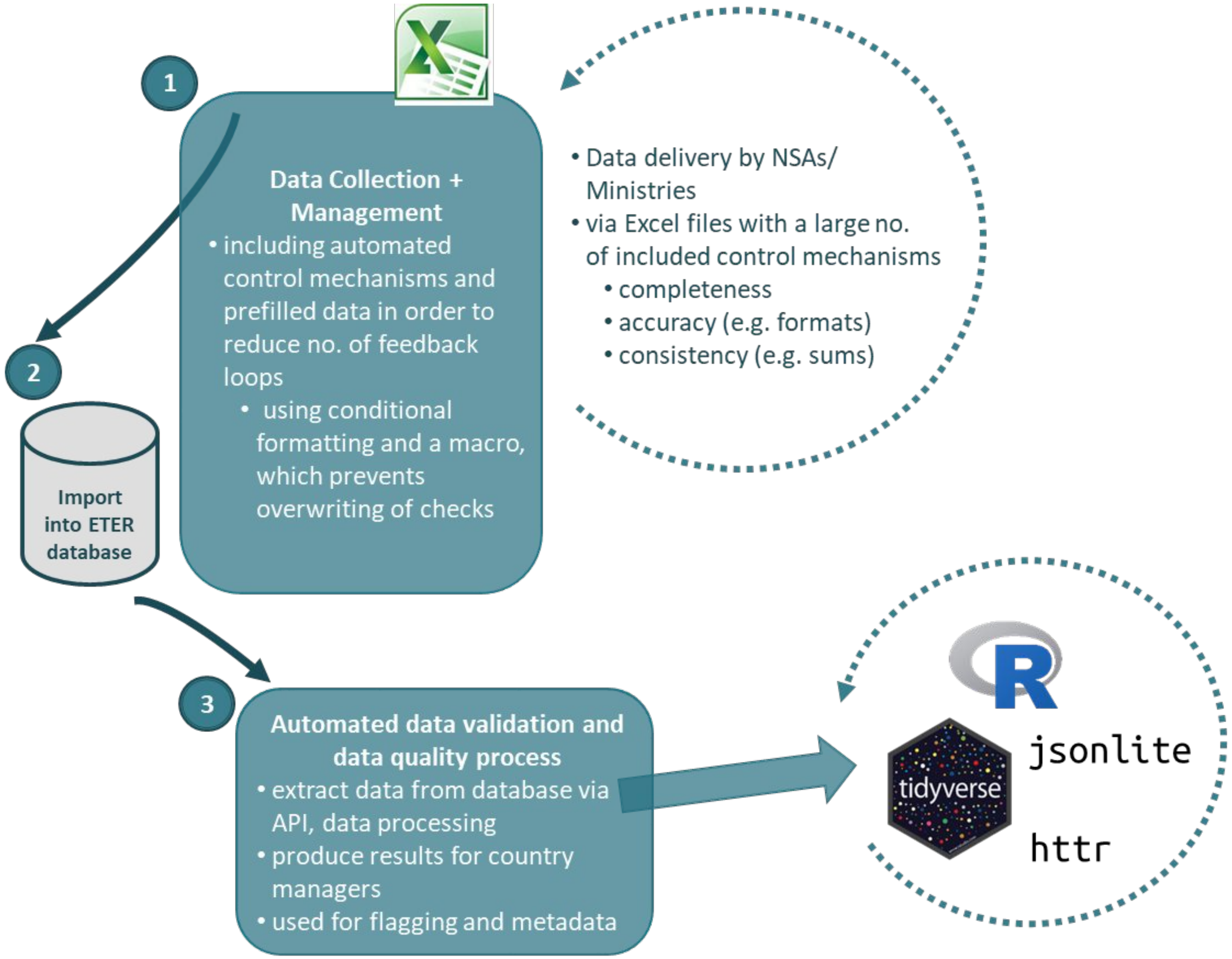
- Collect data from 37+ different sources (countries)

- Ensure tidy data (data validation)
  - Find missing variables, inconsistencies, wrong formats

- Ensure high data quality
  - Multiannual checks
  - Outlier detection
  - Consistency with external data sources (e.g. Eurostat national aggregates)

- Constant interaction between data deliverer, country managers and data collection managers

Challenge #1: Data Collection



1

Data Collection + Management

- including automated control mechanisms and prefilled data in order to reduce no. of feedback loops
    - using conditional formatting and a macro, which prevents overwriting of checks

- Data delivery by NSAs/ Ministries
- via Excel files with a large no. of included control mechanisms
    - completeness
    - accuracy (e.g. formats)
    - consistency (e.g. sums)

2

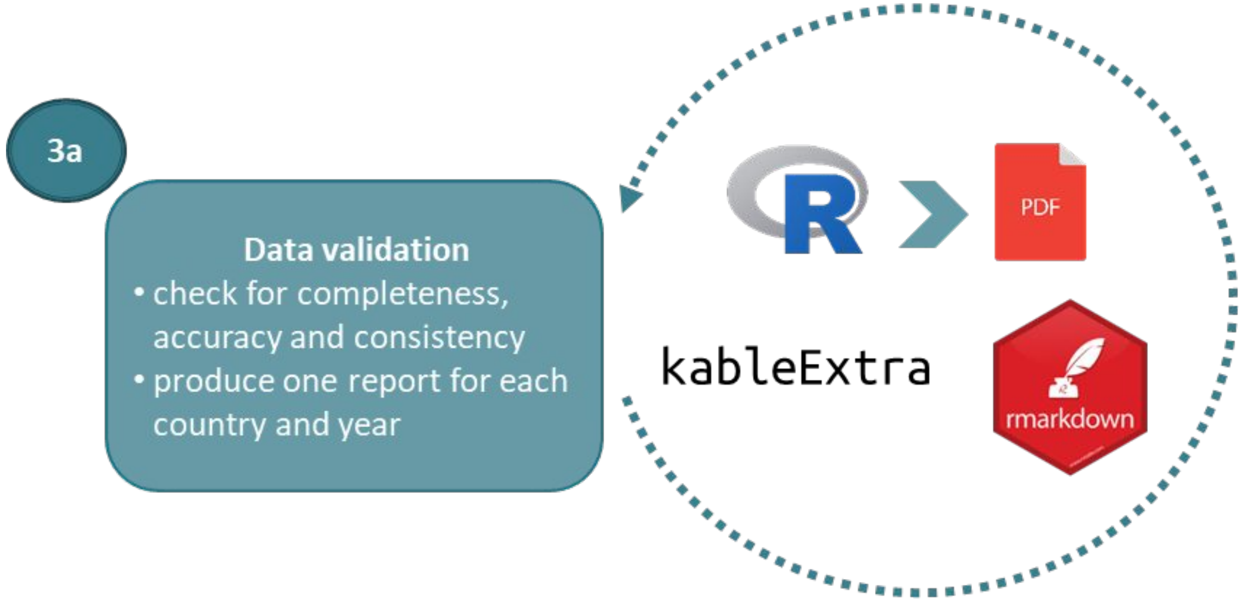
Import into ETER database

3

Automated data validation and data quality process

- extract data from database via API, data processing
- produce results for country managers
- used for flagging and metadata

Challenge #2: Tidy data

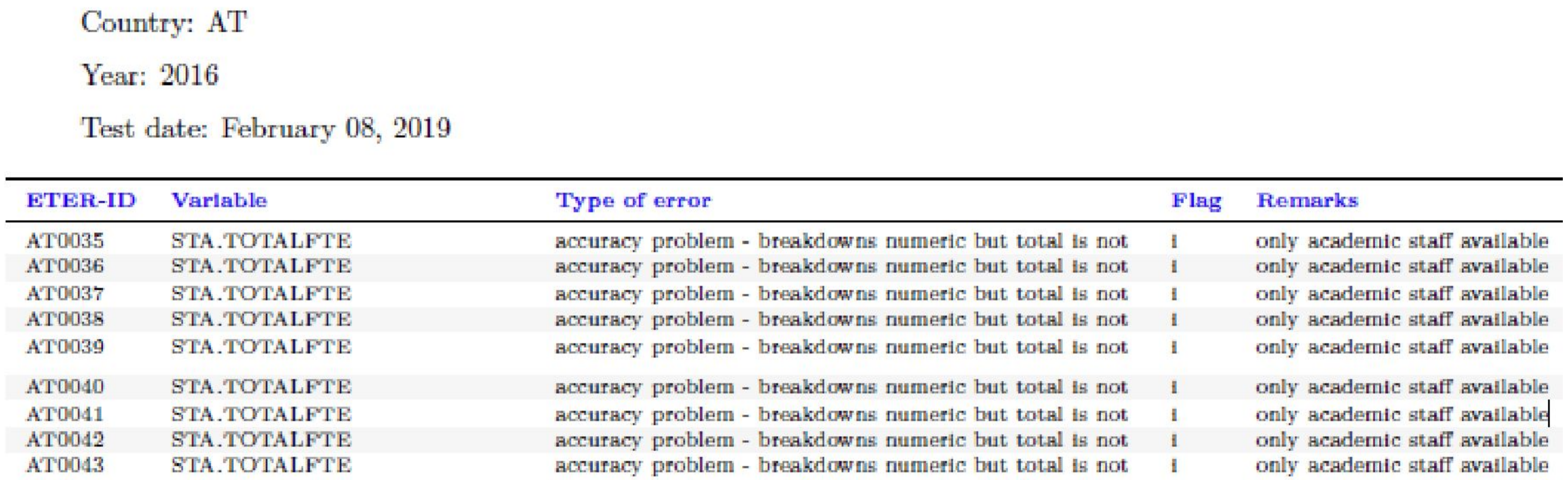


3a

Data validation

- check for completeness, accuracy and consistency
- produce one report for each country and year

Snippet of the Austria data validation report for 2016



Country: AT

Year: 2016

Test date: February 08, 2019

ETER-ID	Variable	Type of error	Flag	Remarks
AT0035	STA.TOTALFTE	accuracy problem - breakdowns numeric but total is not	i	only academic staff available
AT0036	STA.TOTALFTE	accuracy problem - breakdowns numeric but total is not	i	only academic staff available
AT0037	STA.TOTALFTE	accuracy problem - breakdowns numeric but total is not	i	only academic staff available
AT0038	STA.TOTALFTE	accuracy problem - breakdowns numeric but total is not	i	only academic staff available
AT0039	STA.TOTALFTE	accuracy problem - breakdowns numeric but total is not	i	only academic staff available
AT0040	STA.TOTALFTE	accuracy problem - breakdowns numeric but total is not	i	only academic staff available
AT0041	STA.TOTALFTE	accuracy problem - breakdowns numeric but total is not	i	only academic staff available
AT0042	STA.TOTALFTE	accuracy problem - breakdowns numeric but total is not	i	only academic staff available
AT0043	STA.TOTALFTE	accuracy problem - breakdowns numeric but total is not	i	only academic staff available

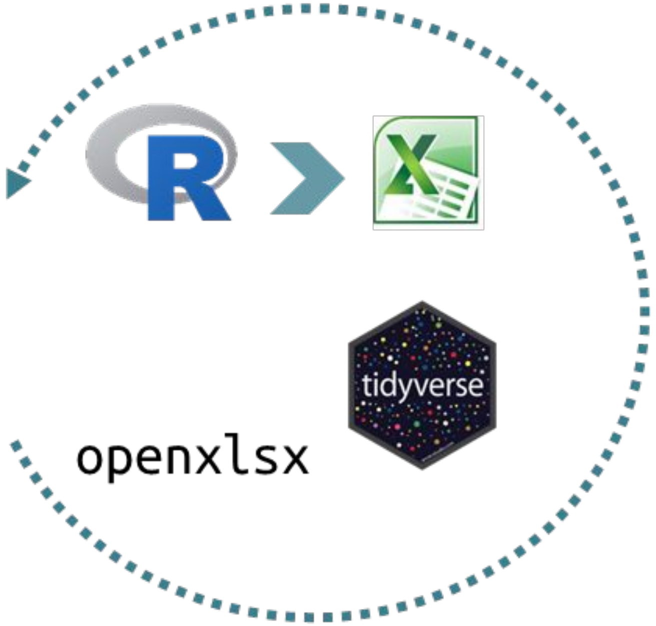
Challenge #3: Data quality

3b

Data quality

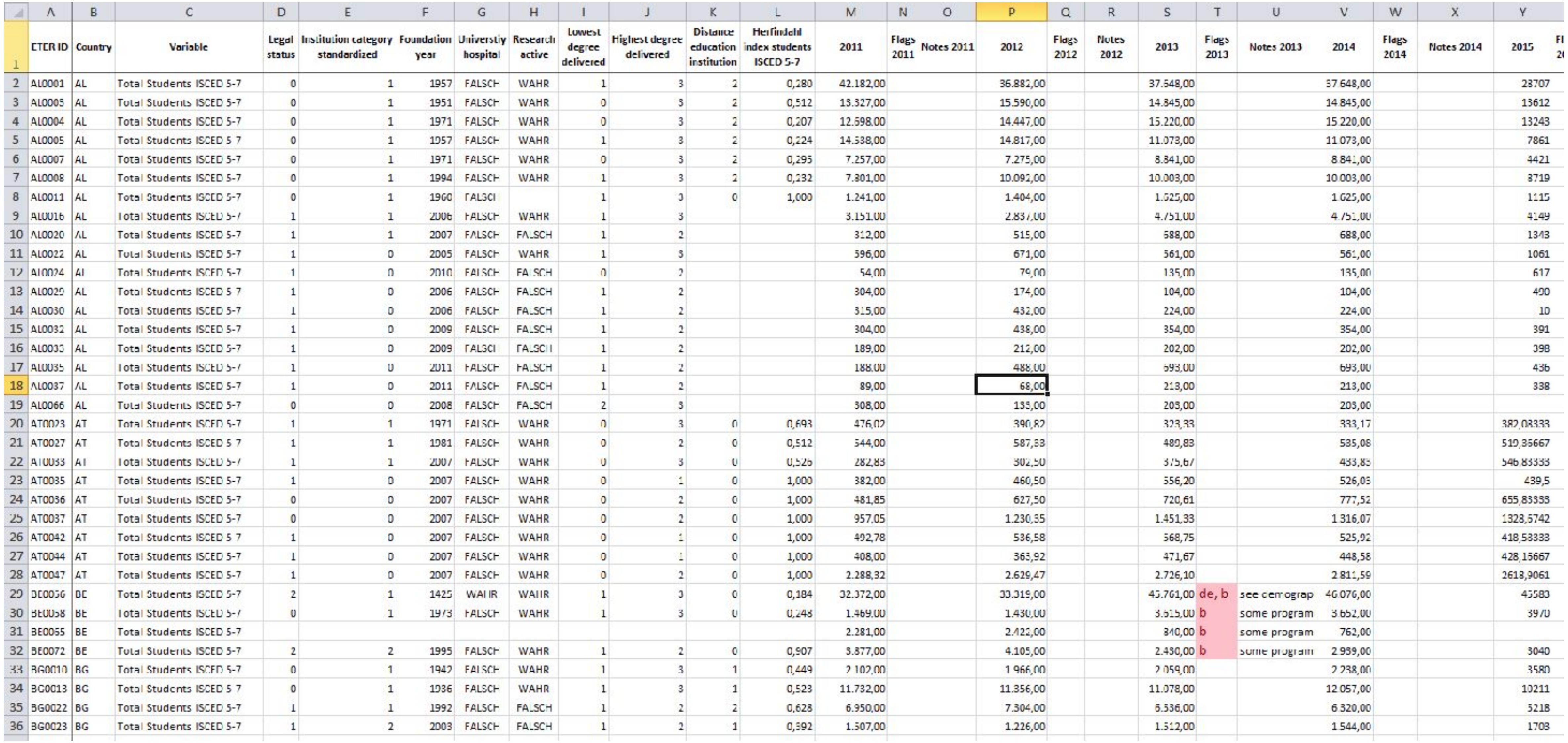
- Internal data quality
  - Multiannual checks
  - Outlier detection
    - determine thresholds
    - calculate test values
    - generate MS-Excel file for country managers
      - use function **conditionalFormatting** of package openxlsx for highlighting already flagged cases

- External data quality
  - e.g. comparison with Eurostat national aggregates



Challenge #3: Data quality (output)

Snippet of the sheet containing positive multiannual test cases



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	
1	ETER ID	Country	Variable	Legal status	Institution category standardized	Foundation year	Universtiy hospital	Research active	Lowest degree delivered	Highest degree delivered	Distance education institution	Herfindahl index students ISCED 5-7	2011	Flags 2011	Notes 2011	2012	Flags 2012	Notes 2012	2013	Flags 2013	Notes 2013	2014	Flags 2014	Notes 2014	2015	Fl 20
2	AL0001	AL	Total Students ISCED 5-7	0	1	1957	FALSCH	WAHR	1	3	2	0,280	42.182,00			36.882,00			37.648,00			37 648,00			23707	
3	AL0003	AL	Total Students ISCED 5-7	0	1	1951	FALSCH	WAHR	0	3	2	0,512	13.327,00			15.590,00			14.845,00			14 845,00			13612	
4	AL0004	AL	Total Students ISCED 5-7	0	1	1971	FALSCH	WAHR	0	3	2	0,207	12.598,00			14.447,00			15.220,00			15 220,00			13243	
5	AL0005	AL	Total Students ISCED 5-7	0	1	1957	FALSCH	WAHR	1	3	2	0,224	14.538,00			14.817,00			11.073,00			11 073,00			7861	
6	AL0007	AL	Total Students ISCED 5-7	0	1	1971	FALSCH	WAHR	0	3	2	0,295	7.257,00			7.275,00			8.841,00			8 841,00			4421	
7	AL0008	AL	Total Students ISCED 5-7	0	1	1994	FALSCH	WAHR	1	3	2	0,232	7.301,00			10.092,00			10.003,00			10 003,00			8719	
8	AL0011	AL	Total Students ISCED 5-7	0	1	1960	FALSCH		1	3	0	1,000	1.241,00			1.404,00			1.625,00			1 625,00			1115	
9	AL0016	AL	Total Students ISCED 5-7	1	1	2006	FALSCH	WAHR	1	3			3.151,00			2.837,00			4.751,00			4 751,00			4149	
10	AL0020	AL	Total Students ISCED 5-7	1	1	2007	FALSCH	FALSCH	1	2			312,00			515,00			588,00			688,00			1343	
11	AL0022	AL	Total Students ISCED 5-7	1	0	2005	FALSCH	WAHR	1	3			596,00			671,00			561,00			561,00			1061	
12	AL0024	AL	Total Students ISCED 5-7	1	0	2010	FALSCH	FALSCH	0	2			54,00			79,00			135,00			135,00			617	
13	AL0029	AL	Total Students ISCED 5-7	1	0	2006	FALSCH	FALSCH	1	2			304,00			174,00			104,00			104,00			400	
14	AL0030	AL	Total Students ISCED 5-7	1	0	2006	FALSCH	FALSCH	1	2			315,00			432,00			224,00			224,00			10	
15	AL0032	AL	Total Students ISCED 5-7	1	0	2009	FALSCH	FALSCH	1	2			304,00			438,00			354,00			354,00			391	
16	AL0033	AL	Total Students ISCED 5-7	1	0	2009	FALSCH	FALSCH	1	2			189,00			212,00			202,00			202,00			398	
17	AL0035	AL	Total Students ISCED 5-7	1	0	2011	FALSCH	FALSCH	1	2			188,00			488,00			693,00			693,00			436	
18	AL0037	AL	Total Students ISCED 5-7	1	0	2011	FALSCH	FALSCH	1	2			89,00			68,00			213,00			213,00			338	
19	AL0066	AL	Total Students ISCED 5-7	0	0	2008	FALSCH	FALSCH	2	3			308,00			135,00			203,00			203,00				
20	AT0023	AT	Total Students ISCED 5-7	1	1	1971	FALSCH	WAHR	0	3	0	0,693	476,02			390,82			323,33			333,17			382,08333	
21	AT0027	AT	Total Students ISCED 5-7	1	1	1981	FALSCH	WAHR	0	2	0	0,512	544,00			587,33			489,83			535,08			519,35667	
22	AT0033	AT	Total Students ISCED 5-7	1	1	2007	FALSCH	WAHR	0	3	0	0,525	282,83			302,50			375,67			433,83			546,83333	
23	AT0035	AT	Total Students ISCED 5-7	1	0	2007	FALSCH	WAHR	0	1	0	1,000	382,00			460,50			556,20			526,03			439,5	
24	AT0036	AT	Total Students ISCED 5-7	0	0	2007	FALSCH	WAHR	0	2	0	1,000	481,85			627,50			720,61			777,52			655,83333	
25	AT0037	AT	Total Students ISCED 5-7	0	0	2007	FALSCH	WAHR	0	2	0	1,000	957,05			1.230,35			1.451,33			1 316,07			1328,6742	
26	AT0042	AT	Total Students ISCED 5-7	1	0	2007	FALSCH	WAHR	0	1	0	1,000	492,78			536,58			568,75			525,92			418,53333	
27	AT0044	AT	Total Students ISCED 5-7	1	0	2007	FALSCH	WAHR	0	1	0	1,000	408,00			363,92			471,67			448,58			428,15667	
28	AT0047	AT	Total Students ISCED 5-7	1	0	2007	FALSCH	WAHR	0	2	0	1,000	2.288,32			2.629,47			2.726,10			2 811,59			2613,9061	
29	BE0056	BE	Total Students ISCED 5-7	2	1	1425	WAHR	WAHR	1	3	0	0,184	32.372,00			33.319,00			45.761,00	de, b	see cerrograp	46 076,00			43583	
30	BE0058	BE	Total Students ISCED 5-7	0	1	1973	FALSCH	WAHR	1	3	0	0,248	1.469,00			1.430,00			3.615,00	b	some program	3 652,00			3970	
31	BE0065	BE	Total Students ISCED 5-7										2.281,00			2.422,00			840,00	b	some program	762,00				
32	BE0072	BE	Total Students ISCED 5-7	2	2	1995	FALSCH	WAHR	1	2	0	0,907	3.877,00			4.105,00			2.430,00	b	some program	2 939,00			3040	
33	BG0010	BG	Total Students ISCED 5-7	0	1	1942	FALSCH	WAHR	1	3	1	0,449	2.102,00			1 966,00			2 059,00			2 238,00			3580	
34	BG0013	BG	Total Students ISCED 5-7	0	1	1936	FALSCH	WAHR	1	3	1	0,523	11.732,00			11.356,00			11.078,00			12 057,00			10211	
35	BG0022	BG	Total Students ISCED 5-7	1	1	1992	FALSCH	FALSCH	1	2	2	0,628	6.950,00			7.304,00			6.536,00			6 320,00			5218	
36	BG0023	BG	Total Students ISCED 5-7	1	2	2003	FALSCH	FALSCH	1	2	1	0,392	1.507,00			1.226,00			1.512,00			1 544,00			1708	

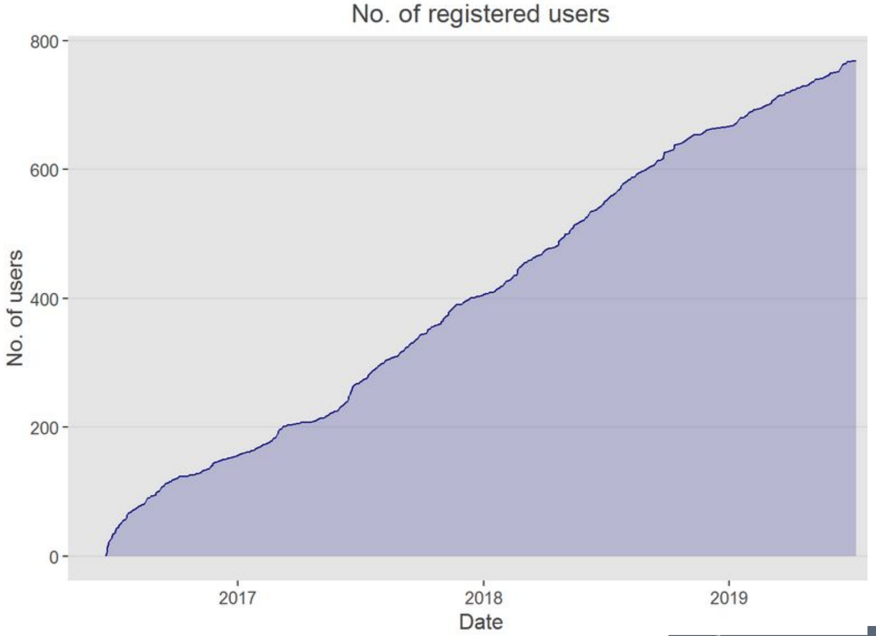
Paper

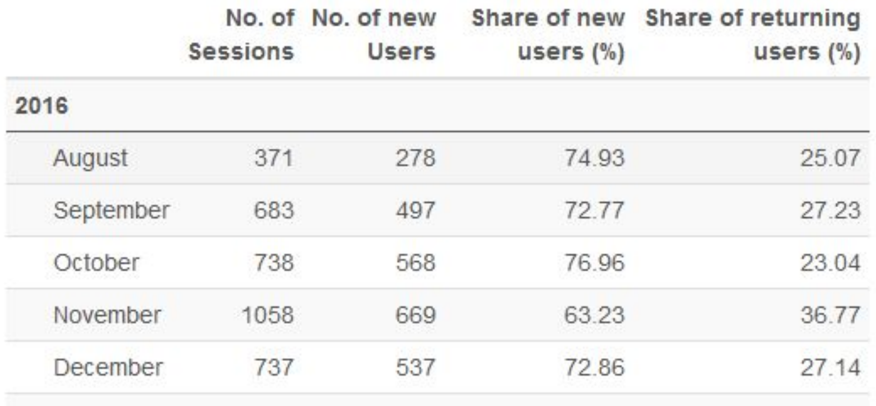
A Tailor-made Data Quality Approach for Higher Educational Data

ETER

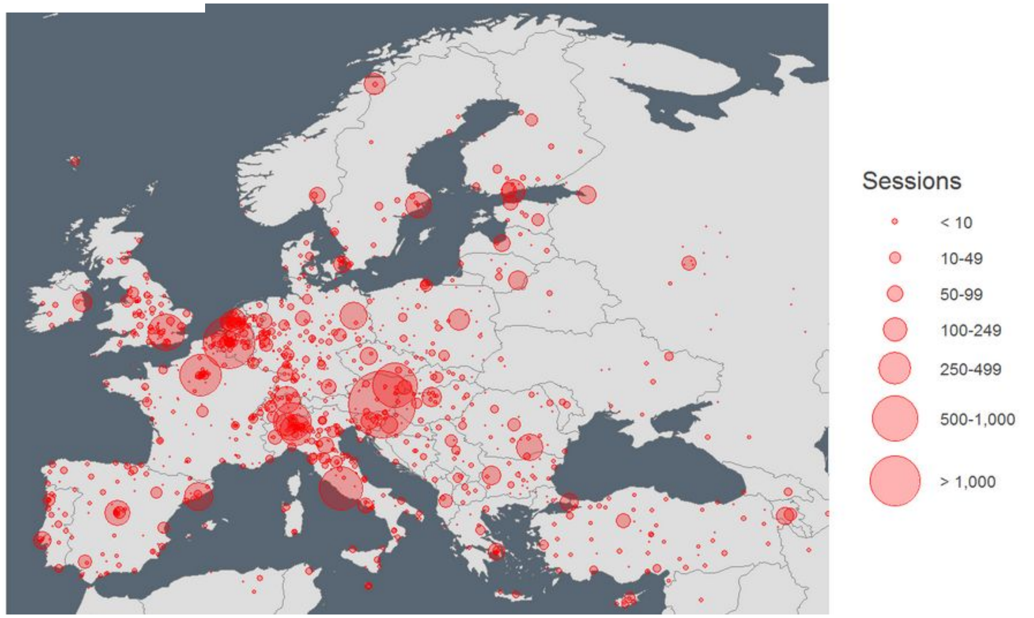
EUROPEAN TERTIARY EDUCATION REGISTER

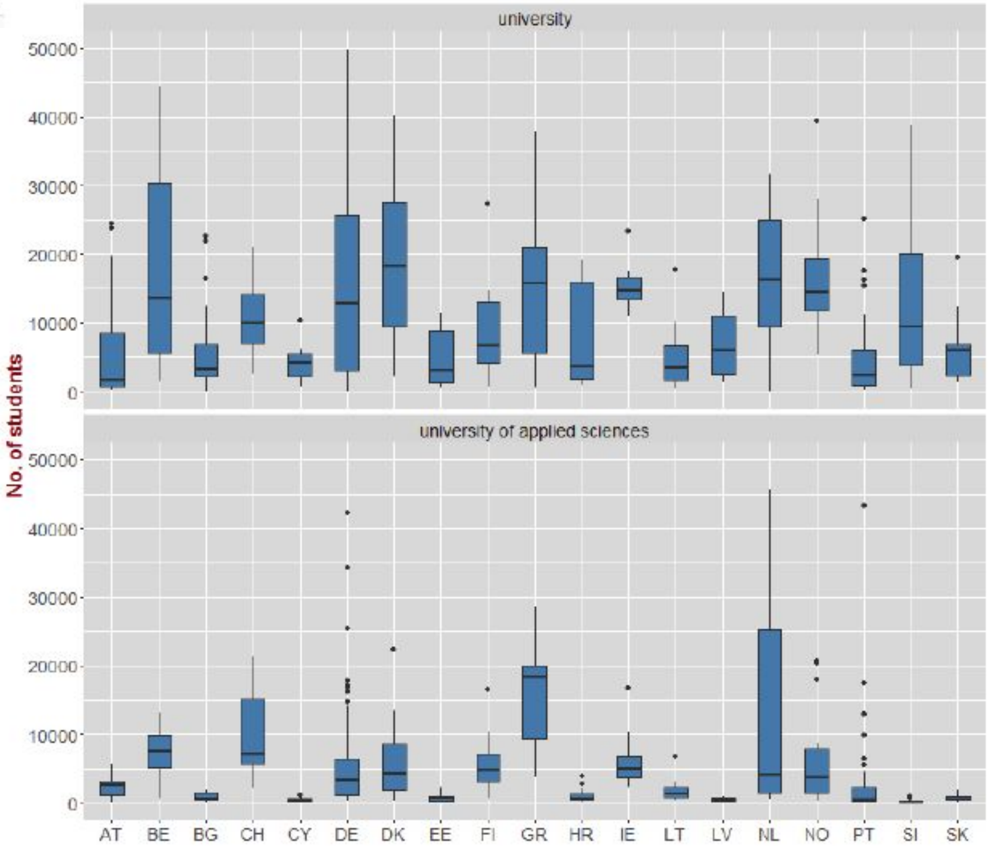
Other R usage in ETER





	No. of Sessions	No. of new Users	Share of new users (%)	Share of returning users (%)
2016				
August	371	278	74.93	25.07
September	683	497	72.77	27.23
October	738	568	76.96	23.04
November	1058	669	63.23	36.77
December	737	537	72.86	27.14





THE INNOVATION COMPANY

JOANNEUM RESEARCH
POLICIES

www.joanneum.at/policies