

Aufgabe 7 Tool Supported Data Cleaning Exercise Carsten Brauer S8963350

1)

#	id	ABC	full_name	ABC	first_name	ABC	last_name	email	gender	#	age
1-21		20 Categories		20 Categories		20 Categories		19 Categories	2 Categories		-78 - 91
	1		Marriel Finnigan		Marriel		Finnigan	mfinnigan@usda.gov	Female		60
	2		Kenyon Possek		Kenyon		Possek	kpossek1@ucoz.com	Male		12
	3		Lalo Manifould		Lalo		Manifould	lmanifould2@pbs.org	Male		26
	4		Nickola Carous		Nickola		Carous	ncarous3@phoca.cz	Male		4
	5		Norman Dubbin		Norman		Dubbin	ndubbin4@wikipedia.org	Male		17
	6		Hasty Perdue		Hasty		Perdue	hperdue5@qq.com			77
	7		Franz Castello		Franz		Castello	fcastello6@1688.com	Male		25
	8		Jorge Tarney		Jorge		Tarney	jtarney7@ft.com			77
	9		Eunice Blakebrough		Eunice		Blakebrough	eblakebrough8@sohu.com	Female		45
	10		Kristopher Frankcombe		Kristopher		Frankcombe	kfrankcombe9@slate.com	Male		old
	11		Palm Domotor		Palm		Domotor	pdomotor@github.io	Male		6
	12		Luz Lansdowne		Luz		Lansdowne	llansdowne@theguardian.com	Female		16
	13		Modestia Keble		Modestia		Keble	mkeblec@cmu.edu	Female		91
	14		Stacee Bovis		Stacee		Bovis	sbovisd@webden.co.uk	Female		22
	15		Eden Wace		Eden		Wace	ewacee@marriott.com	Female		16
	16		Eden Wace		Eden		Wace	ewacee@marriott.com	Female		16
	17		Tobias Sherburn		Tobias		Sherburn	tsherburn@facebook.com	Male		2
											null
	19		Clair Skillern		Clair		Skillern	cskillern@nih.gov	Male		-78
	20		Mathew Addicott		Mathew		Addicott	maddicott@acquirethisname.com	Male		65
	21		Kerianne Goacher		Kerianne		Goacher		Female		45
			Maurits Shawl		Maurits		Shawl	mshawlj@dmz.org	Male		72
											null

Abbildung 1: Rohdaten

#	id	ABC	full_name	ABC	first_name	ABC	last_name	email	gender	#	age
1-20		19 Categories		19 Categories		19 Categories		19 Categories	2 Categories		-78 - 91
	1		Marriel Finnigan		Marriel		Finnigan	mfinnigan@usda.gov	Female		60
	2		Kenyon Possek		Kenyon		Possek	kpossek1@ucoz.com	Male		12
	3		Lalo Manifould		Lalo		Manifould	lmanifould2@pbs.org	Male		26
	4		Nickola Carous		Nickola		Carous	ncarous3@phoca.cz	Male		4
	5		Norman Dubbin		Norman		Dubbin	ndubbin4@wikipedia.org	Male		17
	6		Hasty Perdue		Hasty		Perdue	hperdue5@qq.com			77
	7		Franz Castello		Franz		Castello	fcastello6@1688.com	Male		25
	8		Jorge Tarney		Jorge		Tarney	jtarney7@ft.com	Male		77
	9		Eunice Blakebrough		Eunice		Blakebrough	eblakebrough8@sohu.com	Female		45
	10		Kristopher Frankcombe		Kristopher		Frankcombe	kfrankcombe9@slate.com	Male		old
	11		Palm Domotor		Palm		Domotor	pdomotor@github.io	Male		6
	12		Luz Lansdowne		Luz		Lansdowne	llansdowne@theguardian.com	Female		16
	13		Modestia Keble		Modestia		Keble	mkeblec@cmu.edu	Female		91
	14		Stacee Bovis		Stacee		Bovis	sbovisd@webden.co.uk	Female		22
	15		Eden Wace		Eden		Wace	ewacee@marriott.com	Female		16
	16		Eden Wace		Eden		Wace	ewacee@marriott.com	Female		16
	17		Tobias Sherburn		Tobias		Sherburn	tsherburn@facebook.com	Male		2
	19		Clair Skillern		Clair		Skillern	cskillern@nih.gov	Male		-78
	20		Mathew Addicott		Mathew		Addicott	maddicott@acquirethisname.com	Male		65
			Maurits Shawl		Maurits		Shawl	mshawlj@dmz.org	Male		72

Abbildung 2: filter type: missing missing: email action: Delete

#	id	ABC	full_name	ABC	first_name	ABC	last_name	email	gender	#	age
1-20		18 Categories		18 Categories		18 Categories		18 Categories	2 Categories		-78 - 91
	1		Marriel Finnigan		Marriel		Finnigan	mfinnigan@usda.gov	Female		60
	2		Kenyon Possek		Kenyon		Possek	kpossek1@ucoz.com	Male		12
	3		Lalo Manifould		Lalo		Manifould	lmanifould2@pbs.org	Male		26
	4		Nickola Carous		Nickola		Carous	ncarous3@phoca.cz	Male		4
	5		Norman Dubbin		Norman		Dubbin	ndubbin4@wikipedia.org	Male		17
	6		Hasty Perdue		Hasty		Perdue	hperdue5@qq.com			77
	7		Franz Castello		Franz		Castello	fcastello6@1688.com	Male		25
	8		Jorge Tarney		Jorge		Tarney	jtarney7@ft.com	Male		77
	9		Eunice Blakebrough		Eunice		Blakebrough	eblakebrough8@sohu.com	Female		45
	11		Palm Domotor		Palm		Domotor	pdomotor@github.io	Male		6
	12		Luz Lansdowne		Luz		Lansdowne	llansdowne@theguardian.com	Female		16
	13		Modestia Keble		Modestia		Keble	mkeblec@cmu.edu	Female		91
	14		Stacee Bovis		Stacee		Bovis	sbovisd@webden.co.uk	Female		22
	15		Eden Wace		Eden		Wace	ewacee@marriott.com	Female		16
	16		Eden Wace		Eden		Wace	ewacee@marriott.com	Female		16
	17		Tobias Sherburn		Tobias		Sherburn	tsherburn@facebook.com	Male		2
	19		Clair Skillern		Clair		Skillern	cskillern@nih.gov	Male		-78
	20		Mathew Addicott		Mathew		Addicott	maddicott@acquirethisname.com	Male		65
			Maurits Shawl		Maurits		Shawl	mshawlj@dmz.org	Male		72

Abbildung 3: filter type: mismatched col: age mismatched: Integer action: Delete

Aufgabe 7 Tool Supported Data Cleaning Exercise Carsten Brauer S8963350

#	full_name	first_name	last_name	email	gender	#	age	#	column1
1	Marcel Finnigan	Marcel	Finnigan	mfinnigan@usda.gov	Female				
2	Kenyon Possek	Kenyon	Possek	kpossek1@ucox.com	Male				
3	Lalo Manifould	Lalo	Manifould	lmanifould2@pbs.org	Male				
4	Nickola Carous	Nickola	Carous	ncarous3@phoca.cz	Male				
5	Norman Dubbin	Norman	Dubbin	ndubbin4@wikipedia.org	Male				
6	Hasty Perdue	Hasty	Perdue	hperdue5@qq.com					
7	Franz Castello	Franz	Castello	fcastello6@1688.com	Male				
8	Jorge Tarney	Jorge	Tarney	jtorney7@ft.com	Male				
9	Eunice Blakebrough	Eunice	Blakebrough	eblakebrough8@sohu.com	Female				
11	Palm Domotor	Palm	Domotor	pdomotor9@github.io	Male				
12	Luz Lansdowne	Luz	Lansdowne	llansdowne@theguardian.com	Female				
13	Modestia Keble	Modestia	Keble	mkeblec@cmu.edu	Female				
14	Stacee Bovis	Stacee	Bovis	sbovisd@webden.co.uk	Female				
15	Eden Wace	Eden	Wace	ewacee@marriott.com	Female				
16	Eden Wace	Eden	Wace	ewacee@marriott.com	Female				
17	Tobias Sherburn	Tobias	Sherburn	tsherburnf@facebook.com	Male				
19	Clair Skillern	Clair	Skillern	cskillerng@nih.gov	Male				
20	Mathew Addicott	Mathew	Addicott	maddicott@acquirethisname.com	Male				
	Maurits Shawl	Maurits	Shawl	mshawlj@dmoz.org	Male				

Abbildung 4: derive type: single value: if(age < 0, age * -1, age) as: 'column1'

#	id	full_name	first_name	last_name	email	gender	#	column1
1	20							
2								
3								
4								
5								
6								
7								
8								
9								
11								
12								
13								
14								
15								
16								
17								
19								
20								

Abbildung 5: drop col: age action: Drop

#	id	full_name	first_name	last_name	email	gender	#	age
1	20							
2								
3								
4								
5								
6								
7								
8								
9								
11								
12								
13								
14								
15								
16								
17								
19								
20								

Abbildung 6: rename type: manual mapping: [column1,'age']

Aufgabe 7 Tool Supported Data Cleaning Exercise Carsten Brauer S8963350

#	id	full_name	first_name	last_name	email	gender	age
1-20		18 Categories	18 Categories	18 Categories	18 Categories	2 Categories	2-91
	1	Maríel Finnigan	Maríel	Finnigan	mfinnigan@usda.gov	Female	60
	2	Kenyon Possek	Kenyon	Possek	kpossek1@ucox.com	Male	12
	3	Lalo Manifould	Lalo	Manifould	lmanifould2@pbs.org	Male	26
	4	Nickola Carous	Nickola	Carous	ncarous3@phoca.cz	Male	4
	5	Norman Dubbin	Norman	Dubbin	ndubbin4@wikipedia.org	Male	17
	6	Hasty Perdue	Hasty	Perdue	hperdue5@qq.com	null	77
	7	Franz Castello	Franz	Castello	fcastello6@1688.com	Male	25
	8	Jorge Tarney	Jorge	Tarney	jtorney7@ft.com	Male	77
	9	Eunice Blakebrough	Eunice	Blakebrough	eblakebrough8@sohu.com	Female	45
	11	Palm Domotor	Palm	Domotor	pdomotora9@github.io	Male	6
	12	Luz Lansdowne	Luz	Lansdowne	llansdowneb@theguardian.com	Female	16
	13	Modestia Keble	Modestia	Keble	mkeblec@cmu.edu	Female	91
	14	Stacee Bovis	Stacee	Bovis	sbovisd@webeden.co.uk	Female	22
	15	Eden Wace	Eden	Wace	ewacee@marriott.com	Female	16
	16	Eden Wace	Eden	Wace	ewacee@marriott.com	Female	16
	17	Tobias Sherburn	Tobias	Sherburn	tsherburnf@facebook.com	Male	2
	19	Clair Skillern	Clair	Skillern	cskillerng@nih.gov	Male	78
	20	Mathew Addicott	Mathew	Addicott	maddicott@acquirethisname.com	Male	65
		Maurits Shawl	Maurits	Shawl	mshawlj@dmz.org	Male	72

Abbildung 7: set col: gender value: if(gender == "", null(), \$col)

#	id	full_name	first_name	last_name	email	gender	age
1-21		18 Categories	18 Categories	18 Categories	18 Categories	2 Categories	2-91
	1	Maríel Finnigan	Maríel	Finnigan	mfinnigan@usda.gov	Female	60
	2	Kenyon Possek	Kenyon	Possek	kpossek1@ucox.com	Male	12
	3	Lalo Manifould	Lalo	Manifould	lmanifould2@pbs.org	Male	26
	4	Nickola Carous	Nickola	Carous	ncarous3@phoca.cz	Male	4
	5	Norman Dubbin	Norman	Dubbin	ndubbin4@wikipedia.org	Male	17
	6	Hasty Perdue	Hasty	Perdue	hperdue5@qq.com	null	77
	7	Franz Castello	Franz	Castello	fcastello6@1688.com	Male	25
	8	Jorge Tarney	Jorge	Tarney	jtorney7@ft.com	Male	77
	9	Eunice Blakebrough	Eunice	Blakebrough	eblakebrough8@sohu.com	Female	45
	11	Palm Domotor	Palm	Domotor	pdomotora9@github.io	Male	6
	12	Luz Lansdowne	Luz	Lansdowne	llansdowneb@theguardian.com	Female	16
	13	Modestia Keble	Modestia	Keble	mkeblec@cmu.edu	Female	91
	14	Stacee Bovis	Stacee	Bovis	sbovisd@webeden.co.uk	Female	22
	15	Eden Wace	Eden	Wace	ewacee@marriott.com	Female	16
	16	Eden Wace	Eden	Wace	ewacee@marriott.com	Female	16
	17	Tobias Sherburn	Tobias	Sherburn	tsherburnf@facebook.com	Male	2
	19	Clair Skillern	Clair	Skillern	cskillerng@nih.gov	Male	78
	20	Mathew Addicott	Mathew	Addicott	maddicott@acquirethisname.com	Male	65
	21	Maurits Shawl	Maurits	Shawl	mshawlj@dmz.org	Male	72

Abbildung 8: set col: id value: ifmissing(\$col, max(id) + 1)

full_name	first_name	last_name	email	gender	age
18 Categories	18 Categories	18 Categories	18 Categories	2 Categories	2-91
Maríel Finnigan	Maríel	Finnigan	mfinnigan@usda.gov	Female	60
Kenyon Possek	Kenyon	Possek	kpossek1@ucox.com	Male	12
Lalo Manifould	Lalo	Manifould	lmanifould2@pbs.org	Male	26
Nickola Carous	Nickola	Carous	ncarous3@phoca.cz	Male	4
Norman Dubbin	Norman	Dubbin	ndubbin4@wikipedia.org	Male	17
Hasty Perdue	Hasty	Perdue	hperdue5@qq.com	null	77
Franz Castello	Franz	Castello	fcastello6@1688.com	Male	25
Jorge Tarney	Jorge	Tarney	jtorney7@ft.com	Male	77
Eunice Blakebrough	Eunice	Blakebrough	eblakebrough8@sohu.com	Female	45
Palm Domotor	Palm	Domotor	pdomotora9@github.io	Male	6
Luz Lansdowne	Luz	Lansdowne	llansdowneb@theguardian.com	Female	16
Modestia Keble	Modestia	Keble	mkeblec@cmu.edu	Female	91
Stacee Bovis	Stacee	Bovis	sbovisd@webeden.co.uk	Female	22
Eden Wace	Eden	Wace	ewacee@marriott.com	Female	16
Eden Wace	Eden	Wace	ewacee@marriott.com	Female	16
Tobias Sherburn	Tobias	Sherburn	tsherburnf@facebook.com	Male	2
Clair Skillern	Clair	Skillern	cskillerng@nih.gov	Male	78
Mathew Addicott	Mathew	Addicott	maddicott@acquirethisname.com	Male	65
Maurits Shawl	Maurits	Shawl	mshawlj@dmz.org	Male	72

Abbildung 9: drop col: id action: Drop

Aufgabe 7 Tool Supported Data Cleaning Exercise Carsten Brauer S8963350












ABC	full_name	ABC	first_name	ABC	last_name	✉	email	👤	gender	#	age
											
18 Categories	18 Categories	18 Categories	18 Categories	18 Categories	18 Categories	18 Categories	18 Categories	2 Categories		2 - 91	
Maríel Finnigan	Maríel	Finnigan	mfinnigan@usda.gov	Female		60					
Kenyon Possek	Kenyon	Possek	kpossek1@ucoz.com	Male		12					
Lalo Manifould	Lalo	Manifould	lmanifould2@pbs.org	Male		26					
Nickola Carous	Nickola	Carous	ncarous3@phoca.cz	Male		4					
Norman Dubbin	Norman	Dubbin	ndubbin4@wikipedia.org	Male		17					
Hasty Perdue	Hasty	Perdue	hperdue5@qq.com	null		77					
Franz Castello	Franz	Castello	fcastello6@1688.com	Male		25					
Jorge Tarney	Jorge	Tarney	jtarney7@ft.com	Male		77					
Eunice Blakebrough	Eunice	Blakebrough	eblakebrough8@sohu.com	Female		45					
Palm Domotor	Palm	Domotor	pdomotor9@github.io	Male		6					
Luz Lansdowne	Luz	Lansdowne	llansdowne@theguardian.com	Female		16					
Modestia Keble	Modestia	Keble	mkeblec@cmu.edu	Female		91					
Stacee Bovis	Stacee	Bovis	sbovisd@webeden.co.uk	Female		22					
Eden Wace	Eden	Wace	ewacee@marriott.com	Female		16					
Tobias Sherburn	Tobias	Sherburn	tsherburnf@facebook.com	Male		2					
Clair Skillern	Clair	Skillern	cskillerng@nih.gov	Male		78					
Mathew Addicott	Mathew	Addicott	maddicott@acquirethisname.com	Male		65					
Maurits Shawl	Maurits	Shawl	mshawlj@dmoz.org	Male		72					

Abbildung 10: deduplicate

ABC	full_name	ABC	first_name	ABC	last_name	✉	email	👤	gender	#	age
15 Categories	15 Categories	15 Categories	15 Categories	15 Categories	15 Categories	15 Categories	15 Categories	2 Categories		12 - 91	
Maríel Finnigan	Maríel	Finnigan	mfinnigan@usda.gov	Female		60					
Kenyon Possek	Kenyon	Possek	kpossek1@ucoz.com	Male		12					
Lalo Manifould	Lalo	Manifould	lmanifould2@pbs.org	Male		26					
Norman Dubbin	Norman	Dubbin	ndubbin4@wikipedia.org	Male		17					
Hasty Perdue	Hasty	Perdue	hperdue5@qq.com	null		77					
Franz Castello	Franz	Castello	fcastello6@1688.com	Male		25					
Jorge Tarney	Jorge	Tarney	jtarney7@ft.com	Male		77					
Eunice Blakebrough	Eunice	Blakebrough	eblakebrough8@sohu.com	Female		45					
Luz Lansdowne	Luz	Lansdowne	llansdowne@theguardian.com	Female		16					
Modestia Keble	Modestia	Keble	mkeblec@cmu.edu	Female		91					
Stacee Bovis	Stacee	Bovis	sbovisd@webeden.co.uk	Female		22					
Eden Wace	Eden	Wace	ewacee@marriott.com	Female		16					
Clair Skillern	Clair	Skillern	cskillerng@nih.gov	Male		78					
Mathew Addicott	Mathew	Addicott	maddicott@acquirethisname.com	Male		65					
Maurits Shawl	Maurits	Shawl	mshawlj@dmoz.org	Male		72					

Abbildung 11: filter type: lessThan col: age lessThan: 10 action: Delete

#	id	ABC	full_name	ABC	first_name	ABC	last_name	✉	email	👤	gender	#	age
1 - 15		15 Categories	15 Categories	15 Categories	15 Categories	15 Categories	15 Categories	2 Categories		12 - 91			
	1	Kenyon Possek	Kenyon	Possek	kpossek1@ucoz.com	Male		12					
	2	Eden Wace	Eden	Wace	ewacee@marriott.com	Female		16					
	3	Luz Lansdowne	Luz	Lansdowne	llansdowne@theguardian.com	Female		16					
	4	Norman Dubbin	Norman	Dubbin	ndubbin4@wikipedia.org	Male		17					
	5	Stacee Bovis	Stacee	Bovis	sbovisd@webeden.co.uk	Female		22					
	6	Franz Castello	Franz	Castello	fcastello6@1688.com	Male		25					
	7	Lalo Manifould	Lalo	Manifould	lmanifould2@pbs.org	Male		26					
	8	Eunice Blakebrough	Eunice	Blakebrough	eblakebrough8@sohu.com	Female		45					
	9	Maríel Finnigan	Maríel	Finnigan	mfinnigan@usda.gov	Female		60					
	10	Mathew Addicott	Mathew	Addicott	maddicott@acquirethisname.com	Male		65					
	11	Maurits Shawl	Maurits	Shawl	mshawlj@dmoz.org	Male		72					
	12	Jorge Tarney	Jorge	Tarney	jtarney7@ft.com	Male		77					
	13	Hasty Perdue	Hasty	Perdue	hperdue5@qq.com	null		77					
	14	Clair Skillern	Clair	Skillern	cskillerng@nih.gov	Male		78					
	15	Modestia Keble	Modestia	Keble	mkeblec@cmu.edu	Female		91					

Abbildung 12: derive type: multiple value: rownumber() order: age as: 'id'

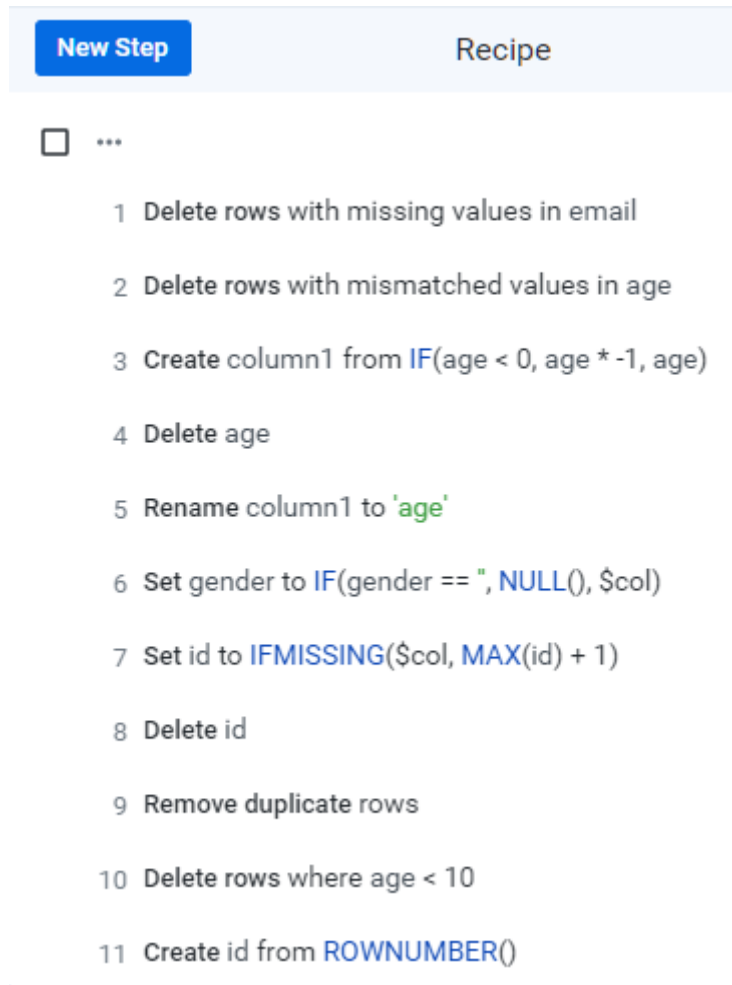


Abbildung 13: Recipe für Aufgabe 1

2)

1. Schwerwiegende Fehler sind in den Spalten, die Zeit oder Datum enthalten. Diese Zeilen würde ich löschen, da es sich dabei nur um wenige Zeilen (67) handelt.
2. Dann lösche ich in allen Text-Spalten die Anführungszeichen.
3. Aus der Spalte „Number of Customers Affected“ entferne ich alle Kommata, wandle den Datentyp in INT um und ersetze alle Mismatches mit NULL
4. Aus der Spalte „Demand Loss (MW)“ entferne ich alle Buchstaben und Sonderzeichen und wandle sie in INT um, nachdem ich alle MISSMATCHED und MISSING in NULL umgewandelt habe.

New Step

Recipe

×

☐ ...

⚙

1

Delete rows with mismatched values in Time Event Began

2

Delete rows with mismatched values in Date of Restoration

3

Delete rows with mismatched values in Time of Restoration

4

Delete rows with missing values in Time of Restoration

5

Replace matches of `"\` from 7 columns with `"`

6

Replace matches of `"` from Number of Customers Affected with `"`

7

Change Number of Customers Affected type to Integer

8

Replace mismatched values in Number of Customers Affected with NULL

9

Replace matches of `{alpha}` from Demand Loss (MW) with `"`

10

Replace matches of `"` from Demand Loss (MW) with `"`

11

Replace matches of `"V` from Demand Loss (MW) with `"`

12

Replace mismatched values in Demand Loss (MW) with NULL

13

Replace missing values in Demand Loss (MW) with NULL

Abbildung 14: Recipe für Aufgabe 2