Data Cleaning Tools

Survey Final

•••

G1

Lukas Bodner, Daniel Geiger, Lorenz Leitner

Introduction

Motivation

Data Sets

Feature Matrix

Motivation

Why is clean data important?

Low-quality data leads to:

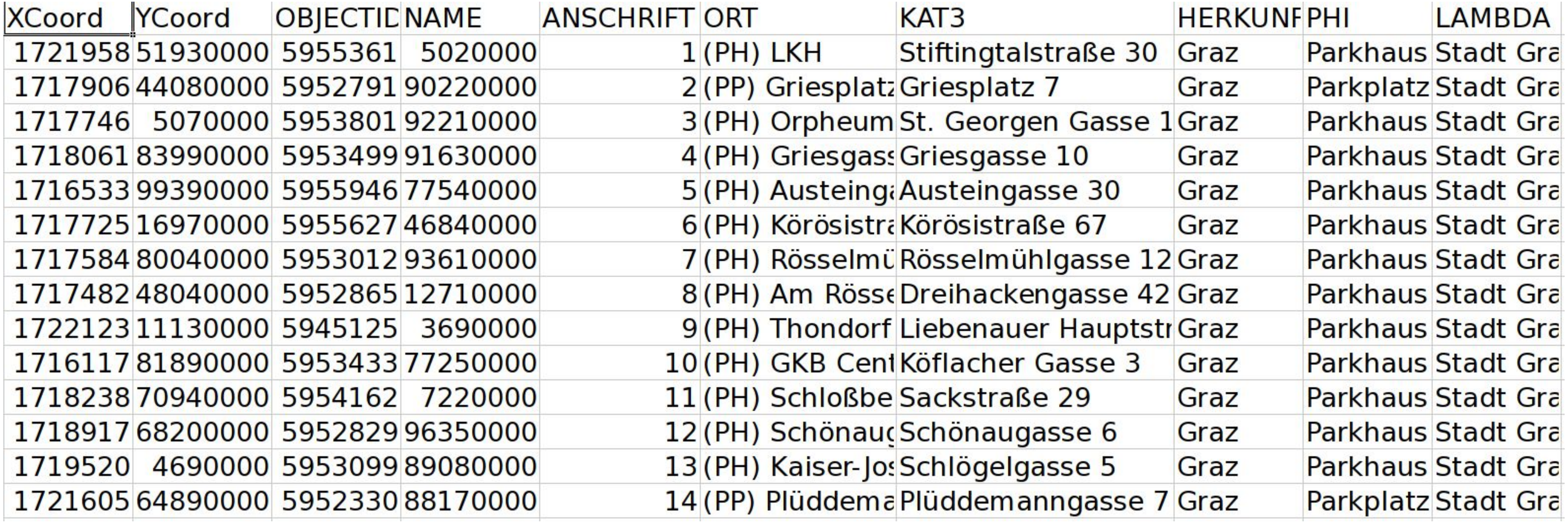
- Incorrect results
- Wrong conclusions

⇒ Costly for businesses

⇒ Material failure

⇒ Injury to people

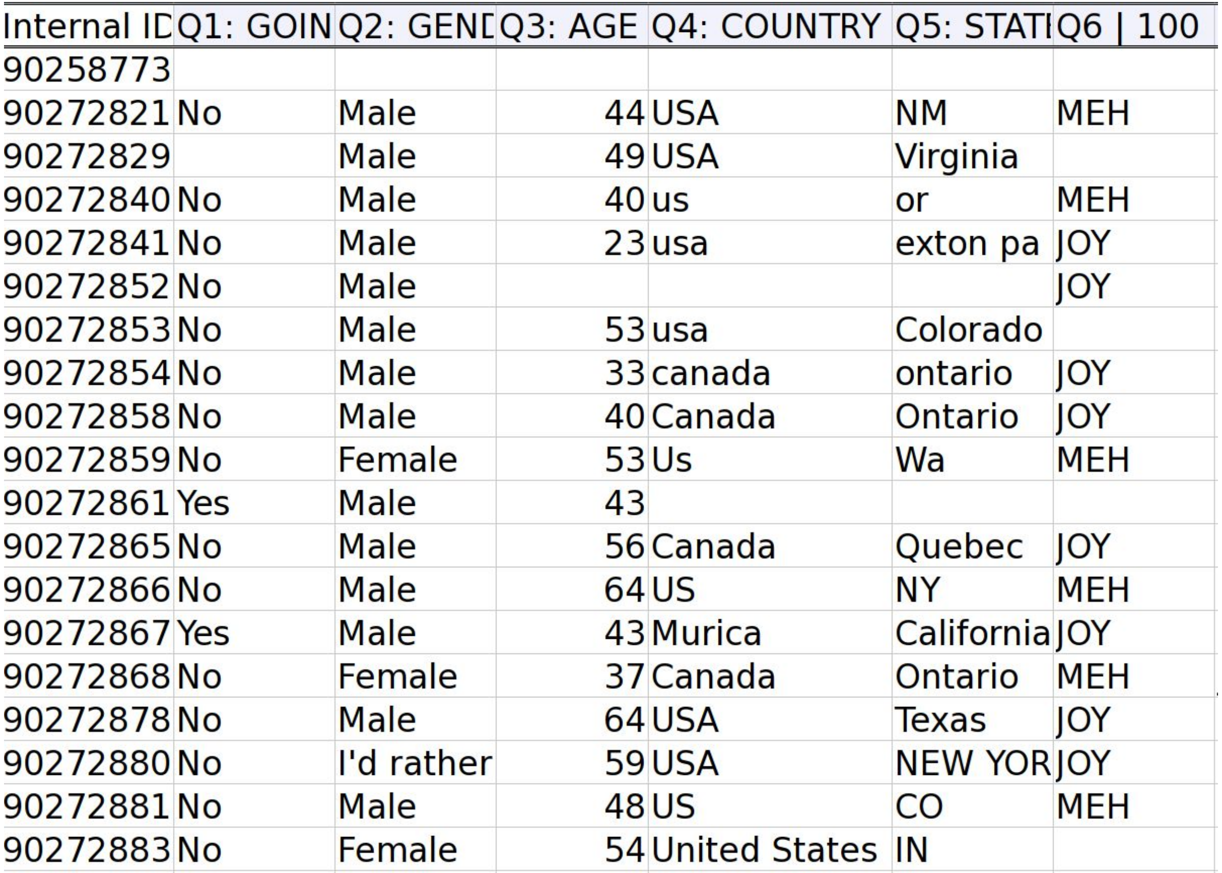
Data Sets - Parking (Task: Merging)



XCoord	YCoord	OBJECTID	NAME	ANSCHRIFT	ORT	KAT3	HERKUNF	PHI	LAMBDA
1721958	51930000	5955361	5020000	1	(PH) LKH	Stiftingtalstraße 30	Graz	Parkhaus	Stadt Gra
1717906	44080000	5952791	90220000	2	(PP) Griesplatz	Griesplatz 7	Graz	Parkplatz	Stadt Gra
1717746	5070000	5953801	92210000	3	(PH) Orpheum	St. Georgen Gasse 1	Graz	Parkhaus	Stadt Gra
1718061	83990000	5953499	91630000	4	(PH) Griesgass	Griesgasse 10	Graz	Parkhaus	Stadt Gra
1716533	99390000	5955946	77540000	5	(PH) Austeinga	Austeingasse 30	Graz	Parkhaus	Stadt Gra
1717725	16970000	5955627	46840000	6	(PH) Körösistra	Körösistraße 67	Graz	Parkhaus	Stadt Gra
1717584	80040000	5953012	93610000	7	(PH) Rösselmü	Rösselmühlgasse 12	Graz	Parkhaus	Stadt Gra
1717482	48040000	5952865	12710000	8	(PH) Am Rösse	Dreihackengasse 42	Graz	Parkhaus	Stadt Gra
1722123	11130000	5945125	3690000	9	(PH) Thondorf	Liebenauer Hauptstr	Graz	Parkhaus	Stadt Gra
1716117	81890000	5953433	77250000	10	(PH) GKB Cent	Köflacher Gasse 3	Graz	Parkhaus	Stadt Gra
1718238	70940000	5954162	7220000	11	(PH) Schloßbe	Sackstraße 29	Graz	Parkhaus	Stadt Gra
1718917	68200000	5952829	96350000	12	(PH) Schönaug	Schönaugasse 6	Graz	Parkhaus	Stadt Gra
1719520	4690000	5953099	89080000	13	(PH) Kaiser-Jos	Schlögelgasse 5	Graz	Parkhaus	Stadt Gra
1721605	64890000	5952330	88170000	14	(PP) Plüddema	Plüddemanngasse 7	Graz	Parkplatz	Stadt Gra

Source:
http://data.graz.gv.at/katalog/verkehr und technik/Parkgaragen.csv
http://data.graz.gv.at/katalog/verkehr und technik/ParkRide.csv

Data Sets - Candy Ratings (Task: Standardization)



Internal ID	Q1: GOIN	Q2: GEND	Q3: AGE	Q4: COUNTRY	Q5: STATI	Q6 | 100
90258773						
90272821	No	Male	44	USA	NM	MEH
90272829		Male	49	USA	Virginia	
90272840	No	Male	40	us	or	MEH
90272841	No	Male	23	usa	exton pa	JOY
90272852	No	Male				JOY
90272853	No	Male	53	usa	Colorado	
90272854	No	Male	33	canada	ontario	JOY
90272858	No	Male	40	Canada	Ontario	JOY
90272859	No	Female	53	Us	Wa	MEH
90272861	Yes	Male	43			
90272865	No	Male	56	Canada	Quebec	JOY
90272866	No	Male	64	US	NY	MEH
90272867	Yes	Male	43	Murica	California	JOY
90272868	No	Female	37	Canada	Ontario	MEH
90272878	No	Male	64	USA	Texas	JOY
90272880	No	I'd rather	59	USA	NEW YOR	JOY
90272881	No	Male	48	US	CO	MEH
90272883	No	Female	54	United States	IN	

Source: https://www.scq.ubc.ca/so-much-candy-data-seriously/

Data Sets - Green Area (Task: Filtering)

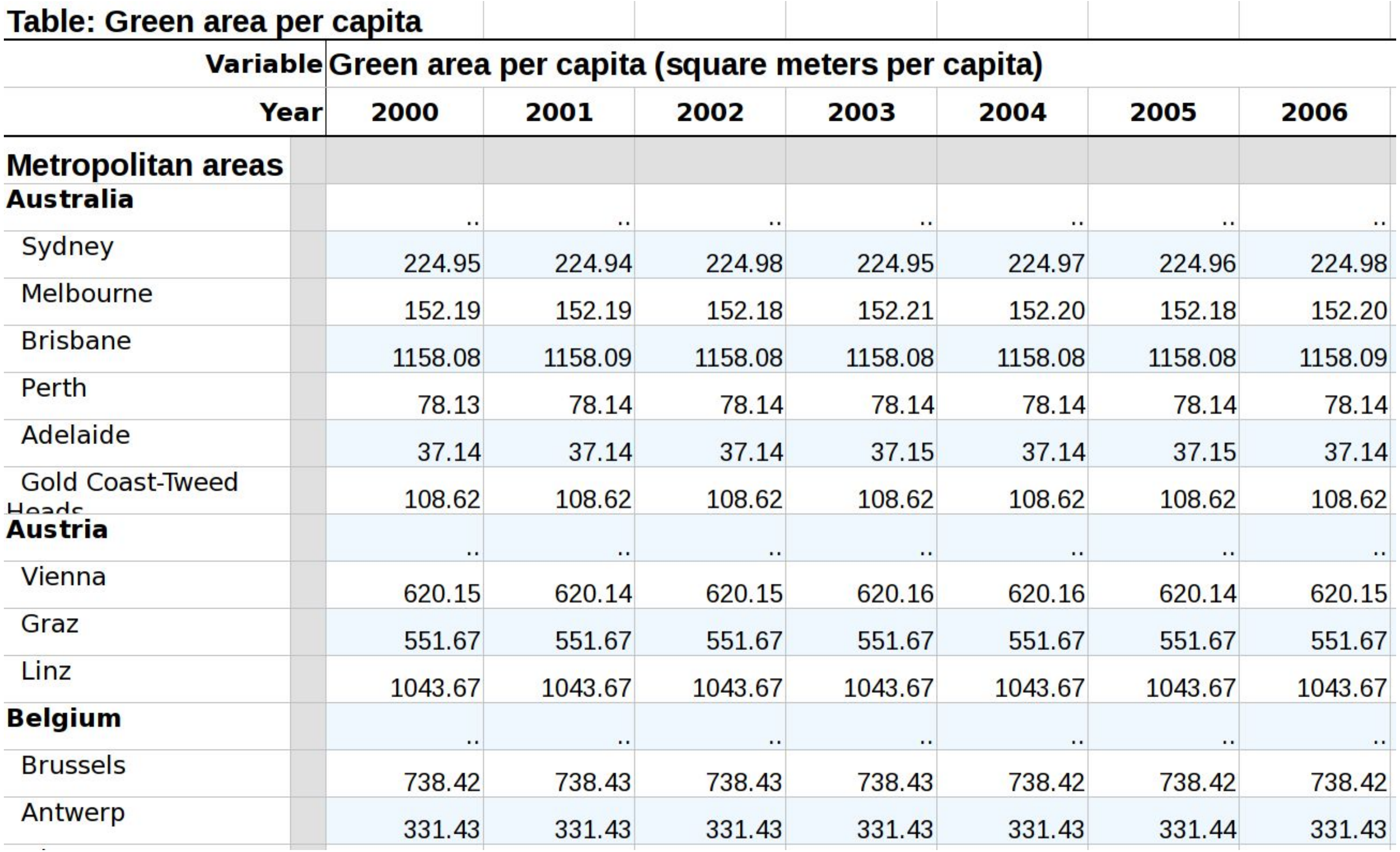


Table: Green area per capita

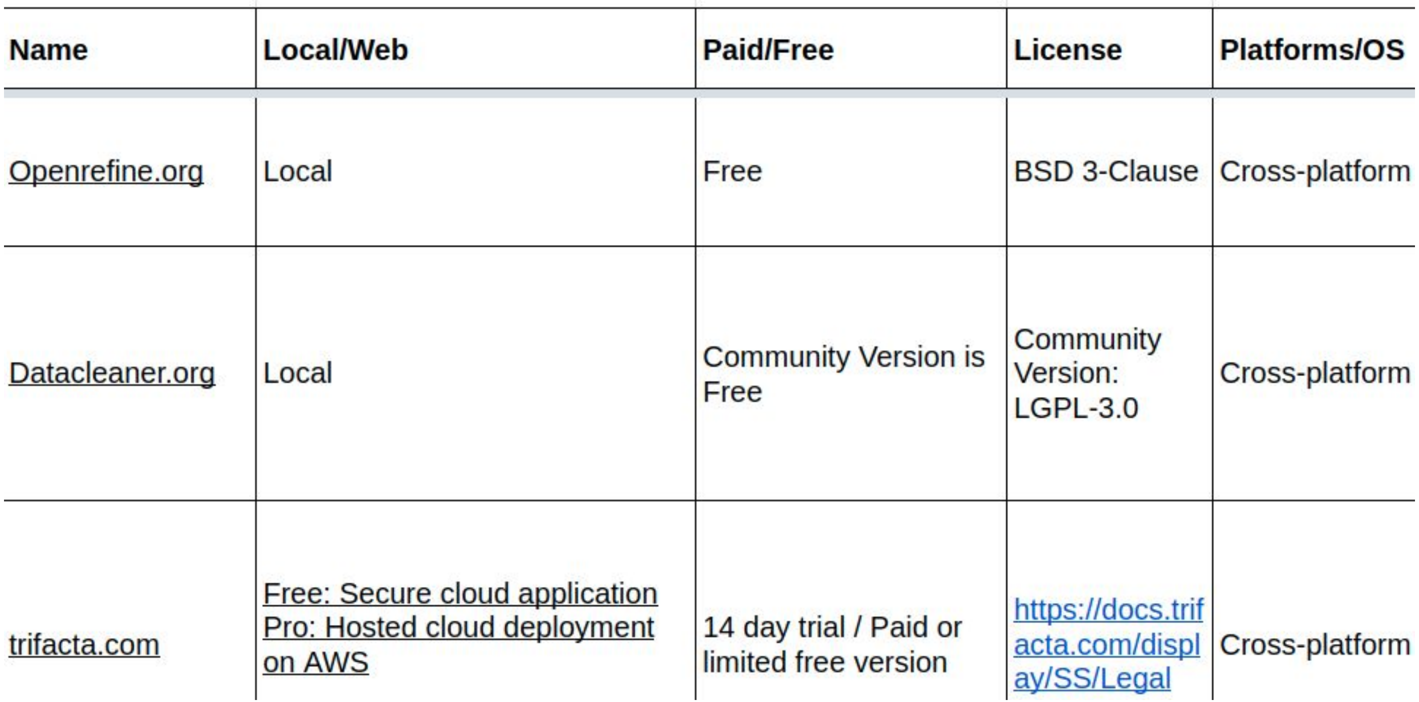
Variable	Green area per capita (square meters per capita)						
Year	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**
Metropolitan areas							
Australia	..	..	..	..	..	..	..
Sydney	224.95	224.94	224.98	224.95	224.97	224.96	224.98
Melbourne	152.19	152.19	152.18	152.21	152.20	152.18	152.20
Brisbane	1158.08	1158.09	1158.08	1158.08	1158.08	1158.08	1158.09
Perth	78.13	78.14	78.14	78.14	78.14	78.14	78.14
Adelaide	37.14	37.14	37.14	37.15	37.14	37.15	37.14
Gold Coast-Tweed Heads	108.62	108.62	108.62	108.62	108.62	108.62	108.62
Austria	..	..	..	..	..	..	..
Vienna	620.15	620.14	620.15	620.16	620.16	620.14	620.15
Graz	551.67	551.67	551.67	551.67	551.67	551.67	551.67
Linz	1043.67	1043.67	1043.67	1043.67	1043.67	1043.67	1043.67
Belgium	..	..	..	..	..	..	..
Brussels	738.42	738.43	738.43	738.43	738.42	738.42	738.42
Antwerp	331.43	331.43	331.43	331.43	331.43	331.44	331.43

Source: https://data.world/unhabitat-guo/7babf915-12a0-4ceb-ad9c-7ee24b776614

Feature Matrix - 12 Characteristics

- Local/Web
- Paid/Free
- License
- Platforms/OS
- Data privacy
- Input formats
- Character encoding
- Output formats
- User-friendliness/ease-of-use
- Documentation
- Support
- Other

Feature Matrix - 25 Tools x 12 Features



Name	Local/Web	Paid/Free	License	Platforms/OS
Openrefine.org	Local	Free	BSD 3-Clause	Cross-platform
Datacleaner.org	Local	Community Version is Free	Community Version: LGPL-3.0	Cross-platform
trifacta.com	Free: Secure cloud application Pro: Hosted cloud deployment on AWS	14 day trial / Paid or limited free version	https://docs.trifacta.com/display/SS/Legal	Cross-platform

Excerpt of the final feature matrix

Tools

Data Cleaning Tools

Descriptions

Evaluations

Examples

OpenRefine

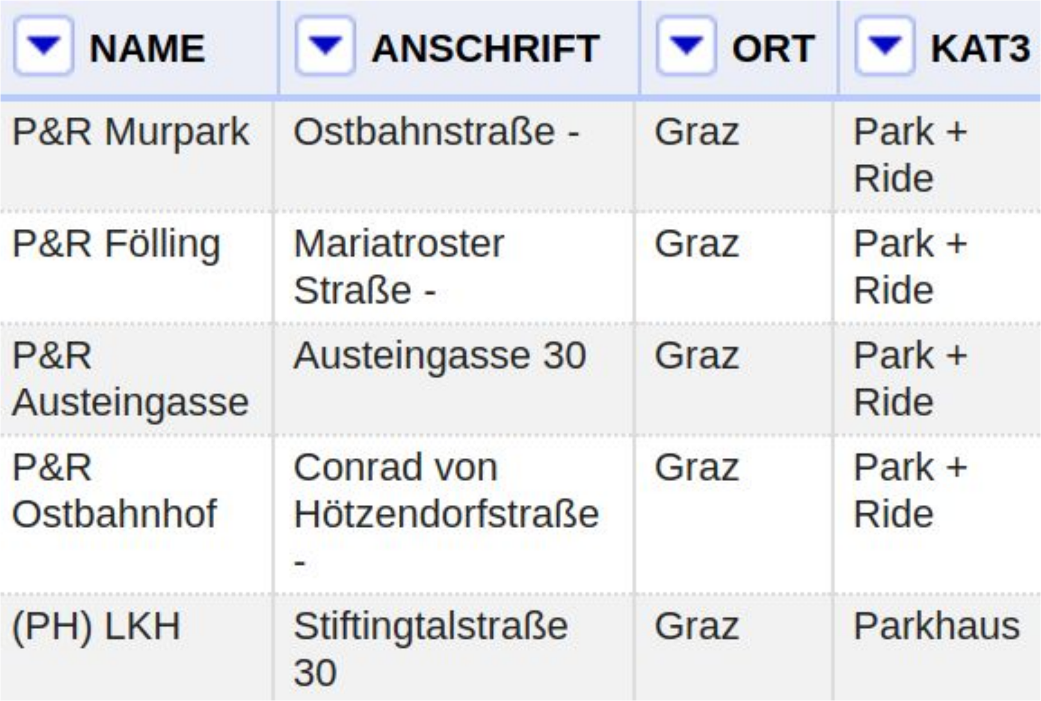
- Local web app
- Free and open source (BSD)
- Cross-platform
- Freebase Gridwork ⇒ Google Refine ⇒ OpenRefine
- Main features
    - Explore data
    - Clean and transform data
    - Match data
    - General Refine Expression Language (GREL)
    - History of applied operations

Showcase video: https://youtu.be/Eqp1OMzW3oQ

OpenRefine - Example 1: Merging

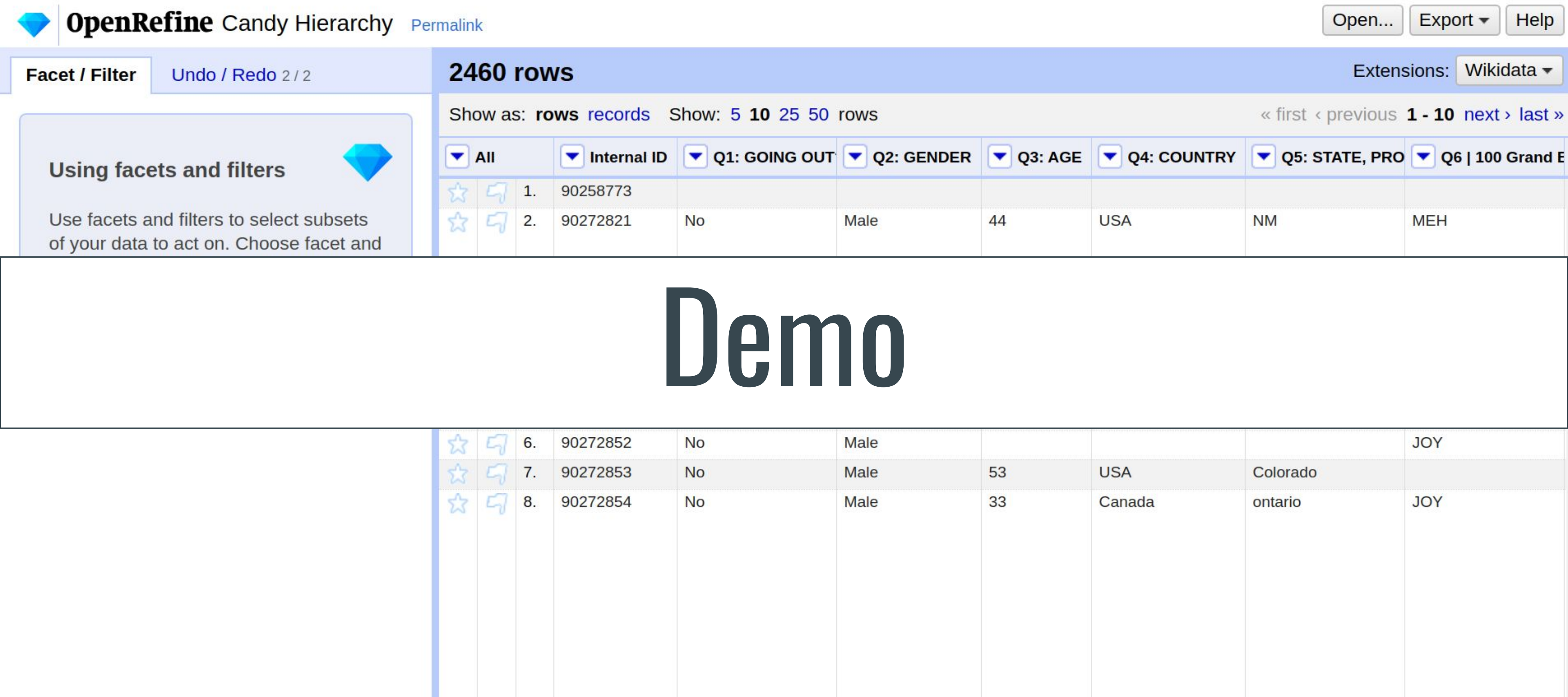
Parkgaragen.csv

ParkRide.csv



▼ NAME	▼ ANSCHRIFT	▼ ORT	▼ KAT3
P&R Murpark	Ostbahnstraße -	Graz	Park + Ride
P&R Fölling	Mariatroster Straße -	Graz	Park + Ride
P&R Austeingasse	Austeingasse 30	Graz	Park + Ride
P&R Ostbahnhof	Conrad von Hötzendorfstraße -	Graz	Park + Ride
(PH) LKH	Stiftingtalstraße 30	Graz	Parkhaus

OpenRefine - Example 2: Standardization



OpenRefine Candy Hierarchy Permalink

Open... Export Help

Facet / Filter Undo / Redo 2 / 2

2460 rows

Extensions: Wikidata

Using facets and filters

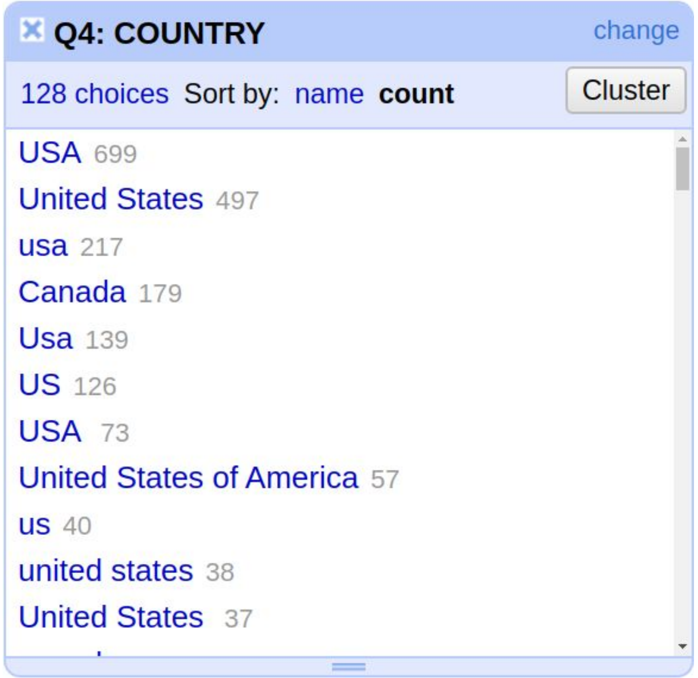
Use facets and filters to select subsets of your data to act on. Choose facet and

Show as: rows records Show: 5 10 25 50 rows « first ‹ previous 1 - 10 next › last »

All	Internal ID	Q1: GOING OUT	Q2: GENDER	Q3: AGE	Q4: COUNTRY	Q5: STATE, PRO	Q6 | 100 Grand E
1.	90258773						
2.	90272821	No	Male	44	USA	NM	MEH
6.	90272852	No	Male				JOY
7.	90272853	No	Male	53	USA	Colorado	
8.	90272854	No	Male	33	Canada	ontario	JOY

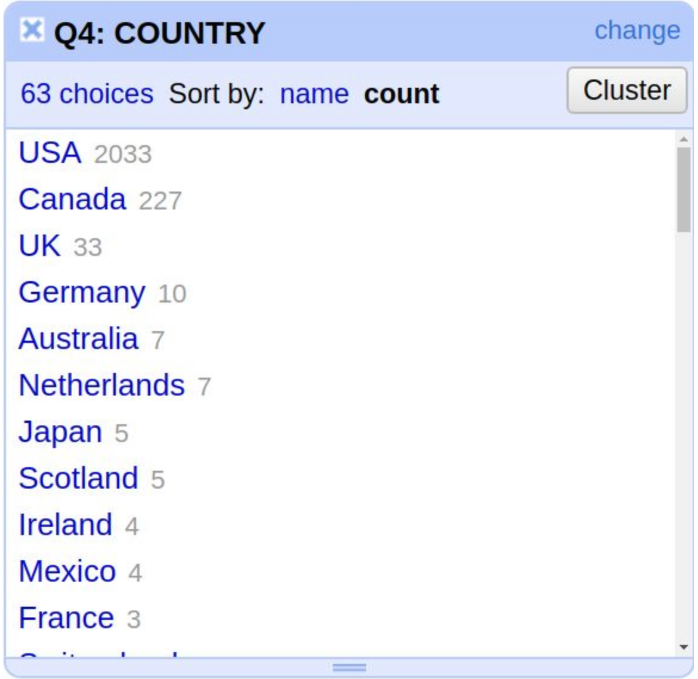
Demo

OpenRefine - Example 2: Standardization



128 choices Sort by: name **count**

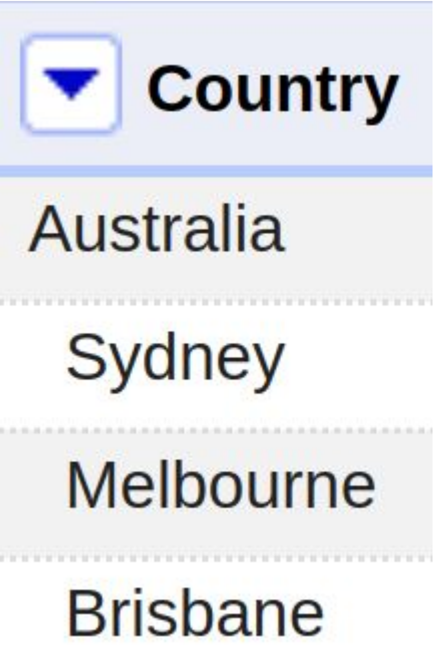
USA	699
United States	497
usa	217
Canada	179
Usa	139
US	126
USA	73
United States of America	57
us	40
united states	38
United States	37



63 choices Sort by: name **count**

USA	2033
Canada	227
UK	33
Germany	10
Australia	7
Netherlands	7
Japan	5
Scotland	5
Ireland	4
Mexico	4
France	3

OpenRefine - Example 3: Filtering (Pre-Processing 1)



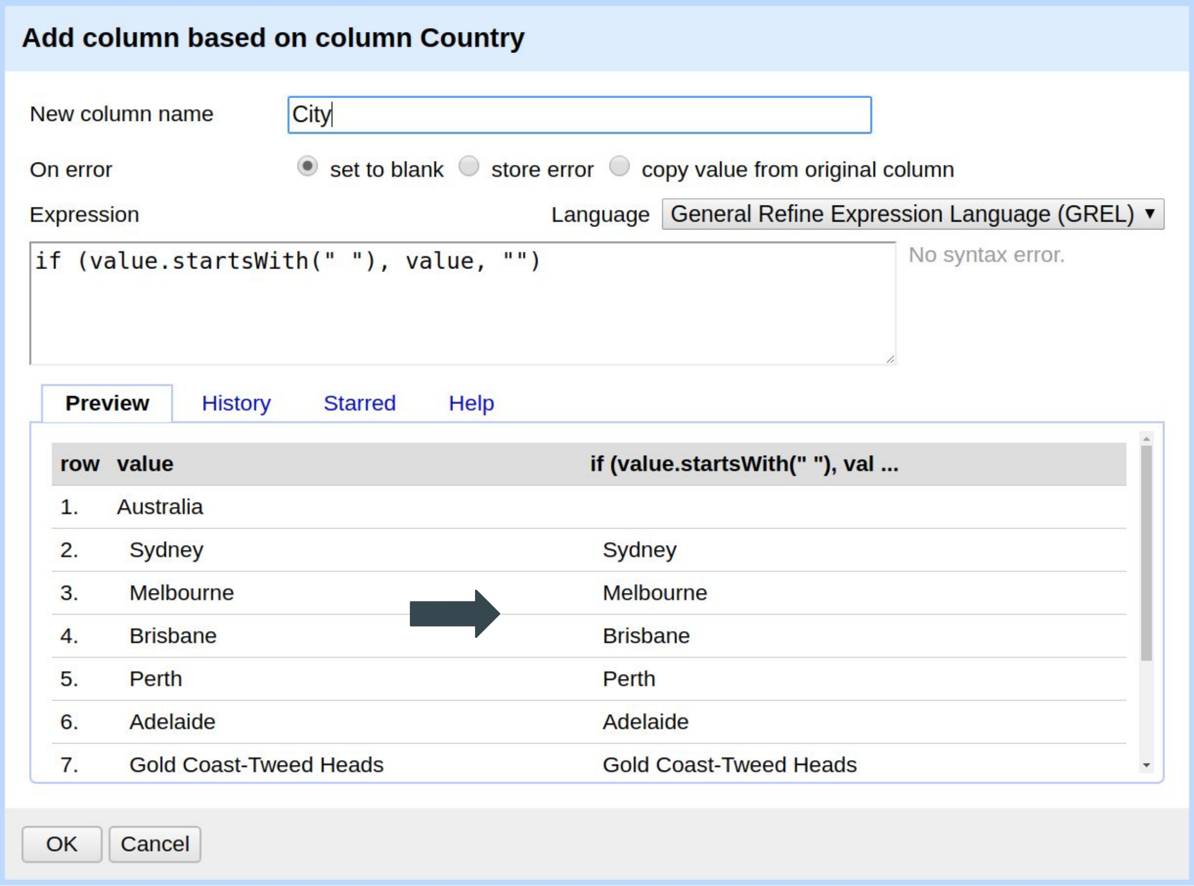
Country

Australia

Sydney

Melbourne

Brisbane



Add column based on column Country

New column name: City

On error: ● set to blank ○ store error ○ copy value from original column

Expression

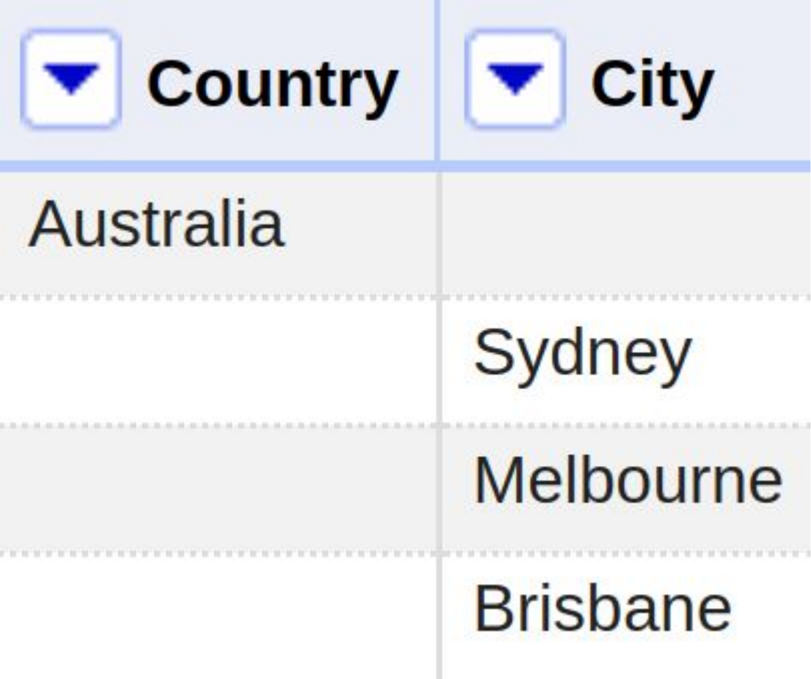
Language: General Refine Expression Language (GREL)

```
if (value.startsWith(" "), value, "")
```

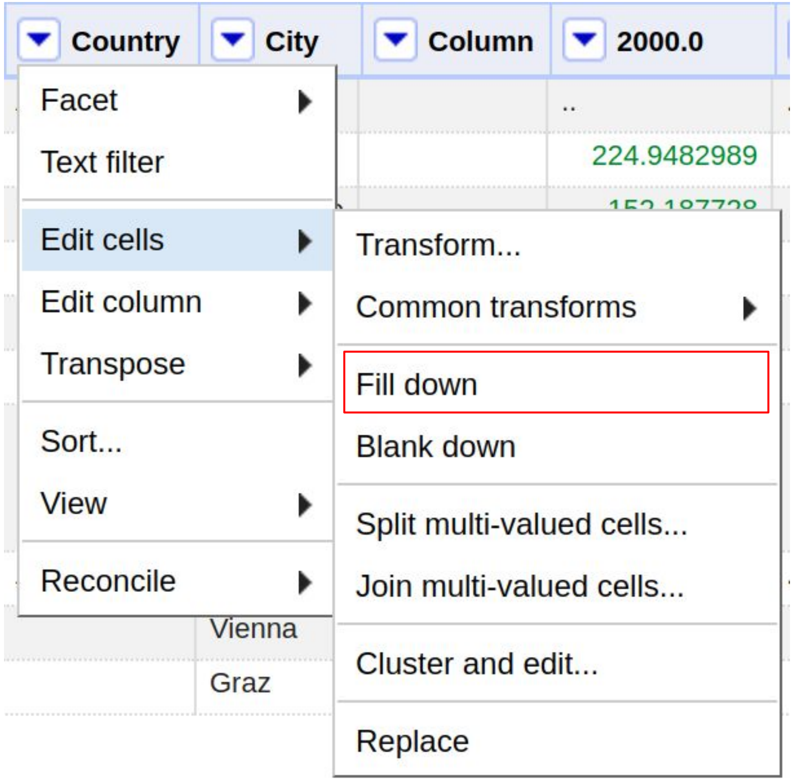
No syntax error.

row	value	if (value.startsWith(" "), val ...
1.	Australia	
2.	Sydney	Sydney
3.	Melbourne	Melbourne
4.	Brisbane	Brisbane
5.	Perth	Perth
6.	Adelaide	Adelaide
7.	Gold Coast-Tweed Heads	Gold Coast-Tweed Heads

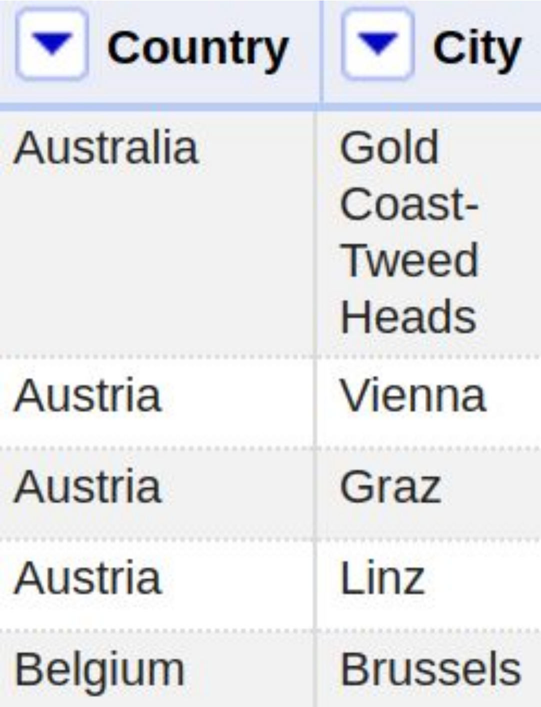
OpenRefine - Example 3: Filtering (Pre-Processing 2)



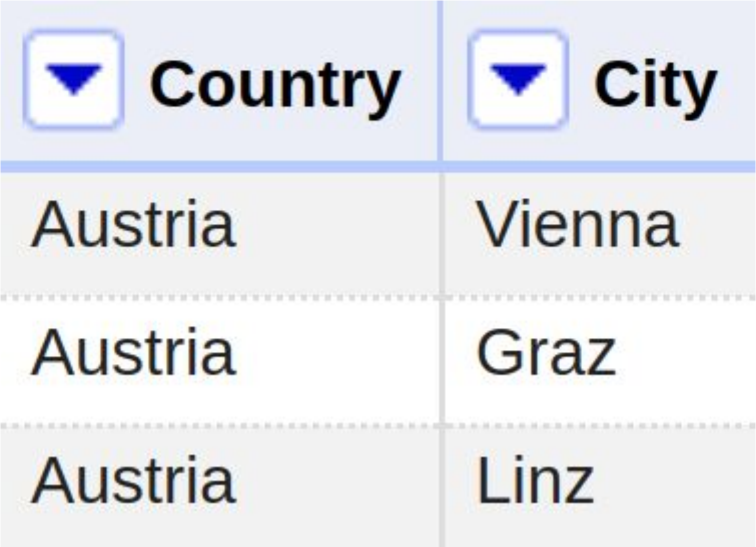
Country	City
Australia	
	Sydney
	Melbourne
	Brisbane



OpenRefine - Example 3: Filtering (Actual Filtering)



Country	City
Australia	Gold Coast-Tweed Heads
Austria	Vienna
Austria	Graz
Austria	Linz
Belgium	Brussels



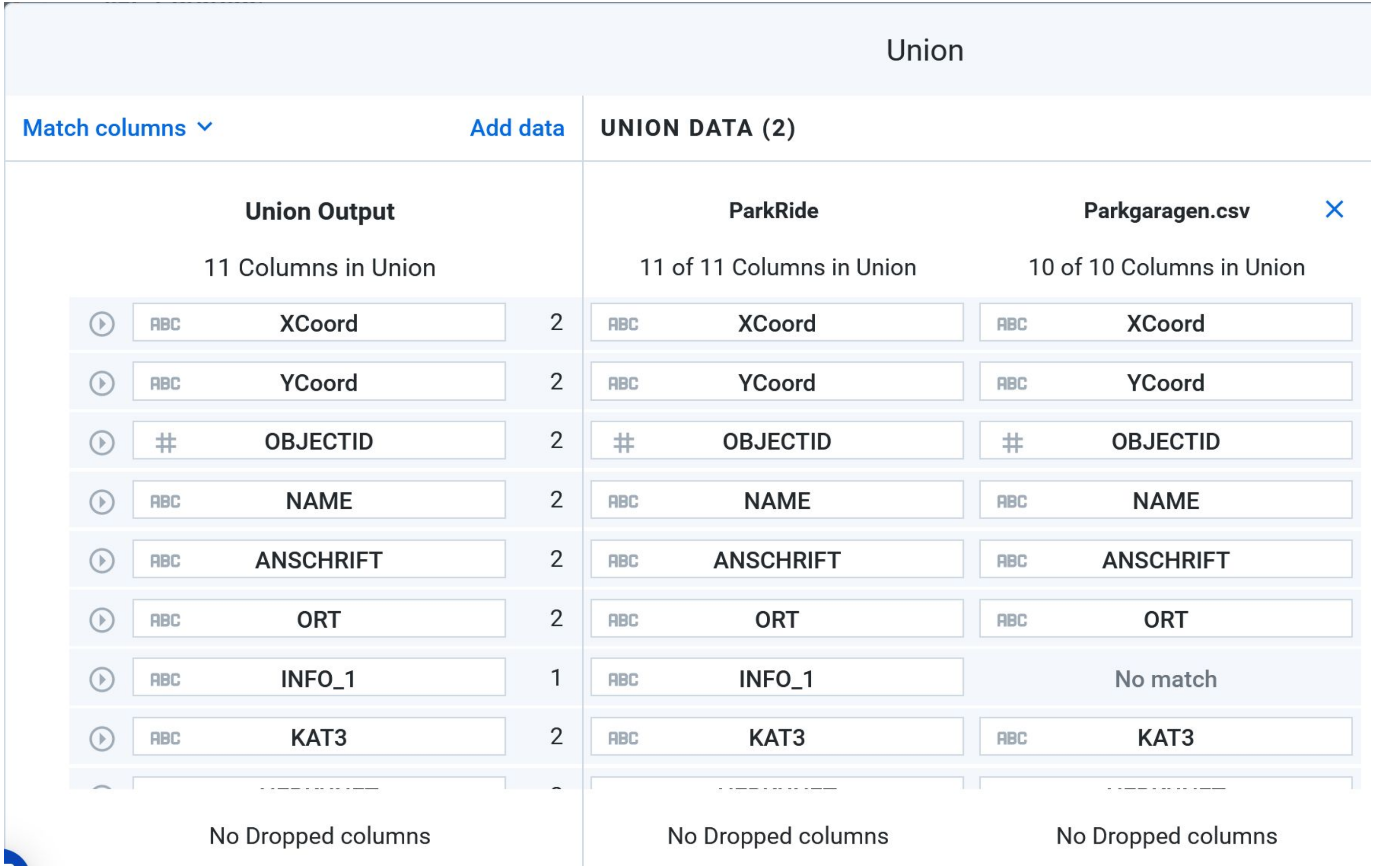
Country	City
Austria	Vienna
Austria	Graz
Austria	Linz

Trifacta

- Web app
- Paid / Free (limited functionality, 100mb upload limit, 1gb download limit)
- Requirements: Chrome and at least 4gb ram (but also works with Firefox)
- Originally called Stanford DataWrangler
- Main features:
  - Suggestions
  - Many transformation functions
  - Preview of transformations
  - Scheduling
- Limitations:
  - Online only

Showcase video: https://youtu.be/HvFGO-U86t8

Trifacta - Example 1: Merging



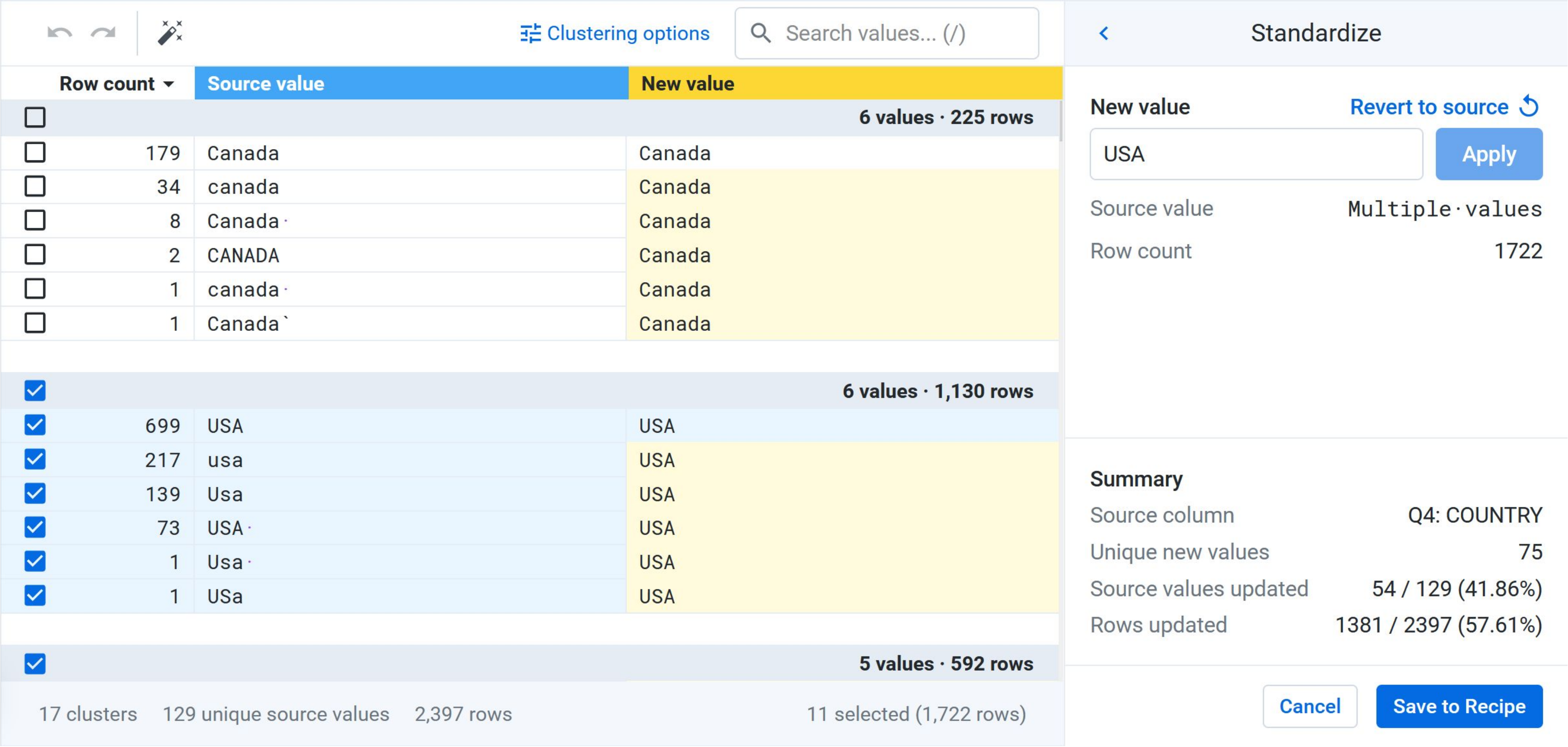
Union

Match columns ⌄ Add data

UNION DATA (2)

Union Output		ParkRide	Parkgaragen.csv ×
11 Columns in Union		11 of 11 Columns in Union	10 of 10 Columns in Union
ABC XCoord	2	ABC XCoord	ABC XCoord
ABC YCoord	2	ABC YCoord	ABC YCoord
# OBJECTID	2	# OBJECTID	# OBJECTID
ABC NAME	2	ABC NAME	ABC NAME
ABC ANSCHRIFT	2	ABC ANSCHRIFT	ABC ANSCHRIFT
ABC ORT	2	ABC ORT	ABC ORT
ABC INFO_1	1	ABC INFO_1	No match
ABC KAT3	2	ABC KAT3	ABC KAT3
No Dropped columns		No Dropped columns	No Dropped columns

Trifacta - Example 2: Standardization



	Row count	Source value	New value
□			6 values · 225 rows
□	179	Canada	Canada
□	34	canada	Canada
□	8	Canada·	Canada
□	2	CANADA	Canada
□	1	canada·	Canada
□	1	Canada`	Canada
☑			6 values · 1,130 rows
☑	699	USA	USA
☑	217	usa	USA
☑	139	Usa	USA
☑	73	USA·	USA
☑	1	Usa·	USA
☑	1	USa	USA
☑			5 values · 592 rows

17 clusters 129 unique source values 2,397 rows 11 selected (1,722 rows)

Standardize

New value Revert to source

USA Apply

Source value	Multiple·values
Row count	1722

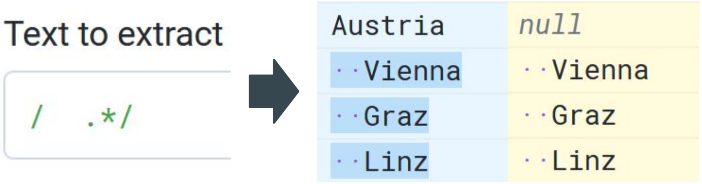
Summary

Source column	Q4: COUNTRY
Unique new values	75
Source values updated	54 / 129 (41.86%)
Rows updated	1381 / 2397 (57.61%)

Cancel Save to Recipe

Trifacta - Example 3: Filtering (Pre-Processing)

1.

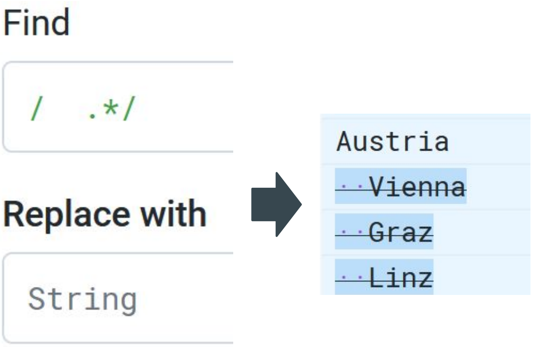


Text to extract

```
/  .*/
```

Austria	*null*
··Vienna	··Vienna
··Graz	··Graz
··Linz	··Linz

2.



Find

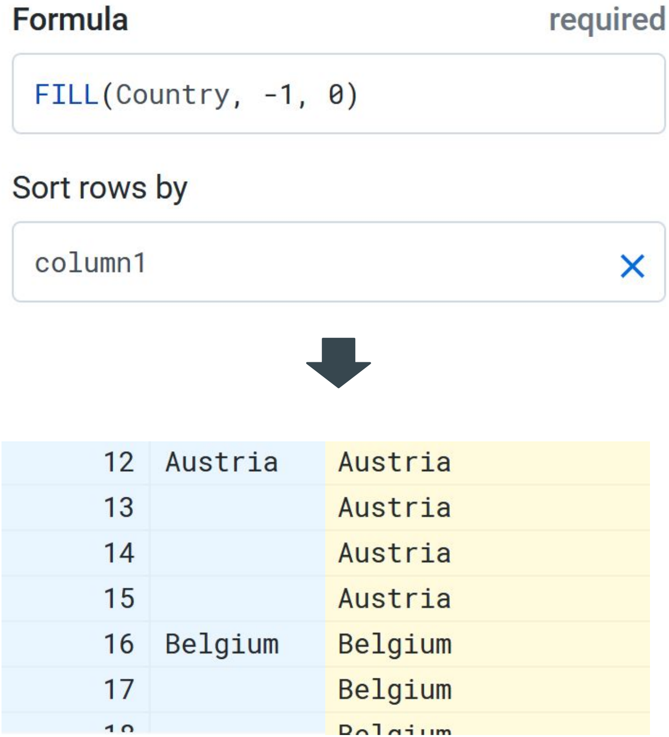
```
/  .*/
```

Replace with

String

Austria
~~··Vienna~~
~~··Graz~~
~~··Linz~~

3.



Formula required

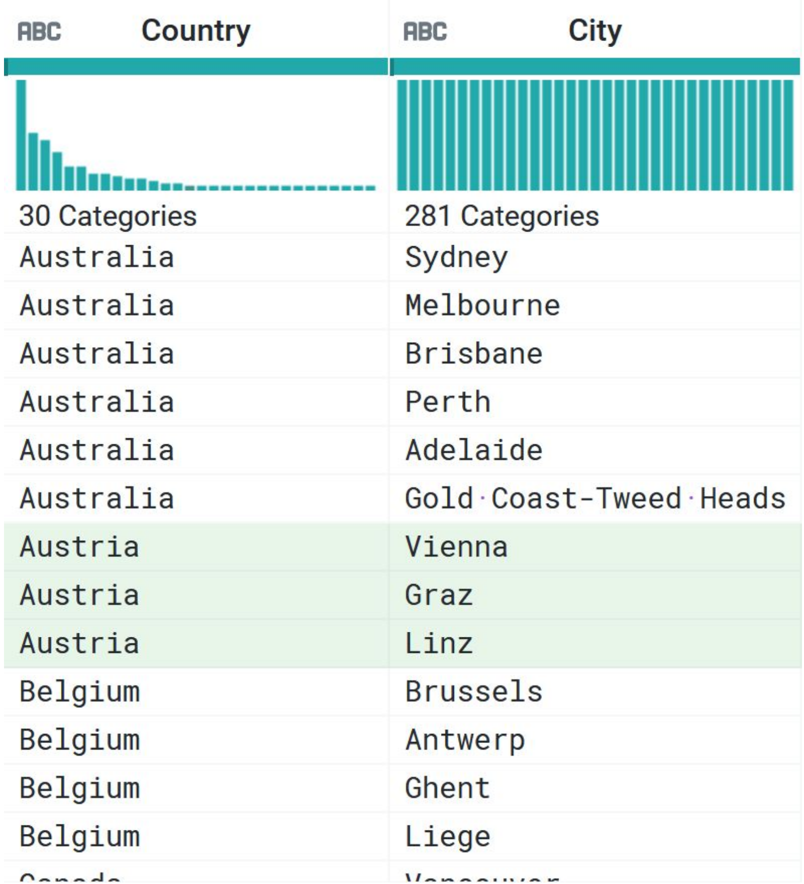
```
FILL(Country, -1, 0)
```

Sort rows by

column1 ✕

12	Austria	Austria
13		Austria
14		Austria
15		Austria
16	Belgium	Belgium
17		Belgium
18		Belgium

Trifacta - Example 3: Filtering (Actual Filtering)



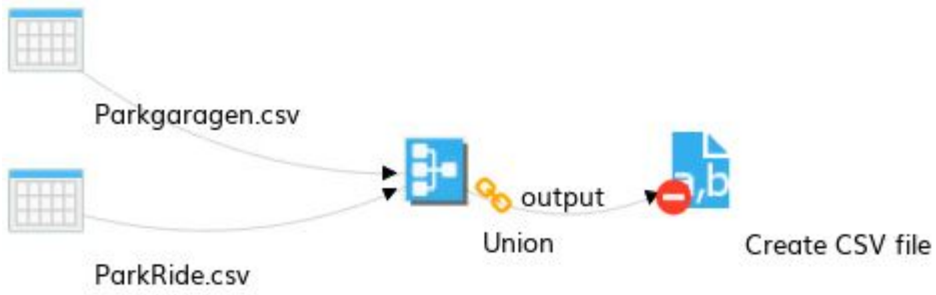
ABC Country	ABC City
30 Categories	281 Categories
Australia	Sydney
Australia	Melbourne
Australia	Brisbane
Australia	Perth
Australia	Adelaide
Australia	Gold·Coast-Tweed·Heads
Austria	Vienna
Austria	Graz
Austria	Linz
Belgium	Brussels
Belgium	Antwerp
Belgium	Ghent
Belgium	Liege
Canada	Vancouver

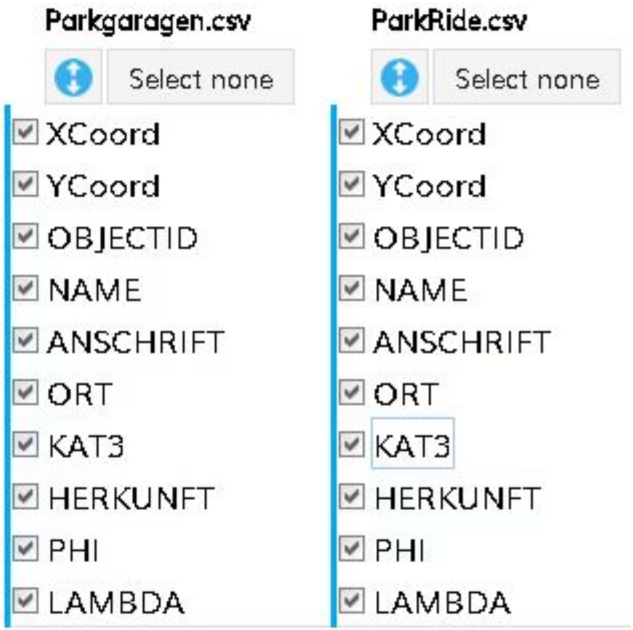
DataCleaner

- Standalone desktop application
- Paid commercial edition and free and open-source community edition (LGPL-3.0)
- Cross-platform
- First released in 2008
- Main features:
    - Data profiling (Discovering and analyzing quality of data)
    - Data wrangling (Transforming and cleaning data)
    - Community driven extensions
- Limitations:
    - In practice many errors and crashes
    - Unintuitive usage

Showcase video: https://youtu.be/bvLEYrTC6CY

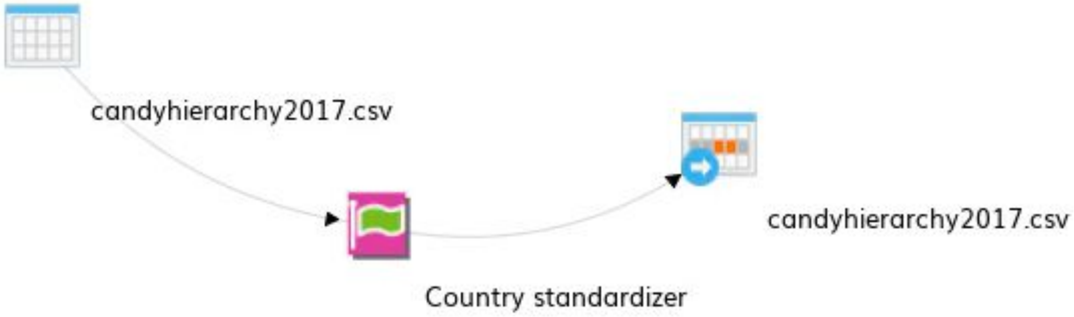
DataCleaner - Example 1: Merging

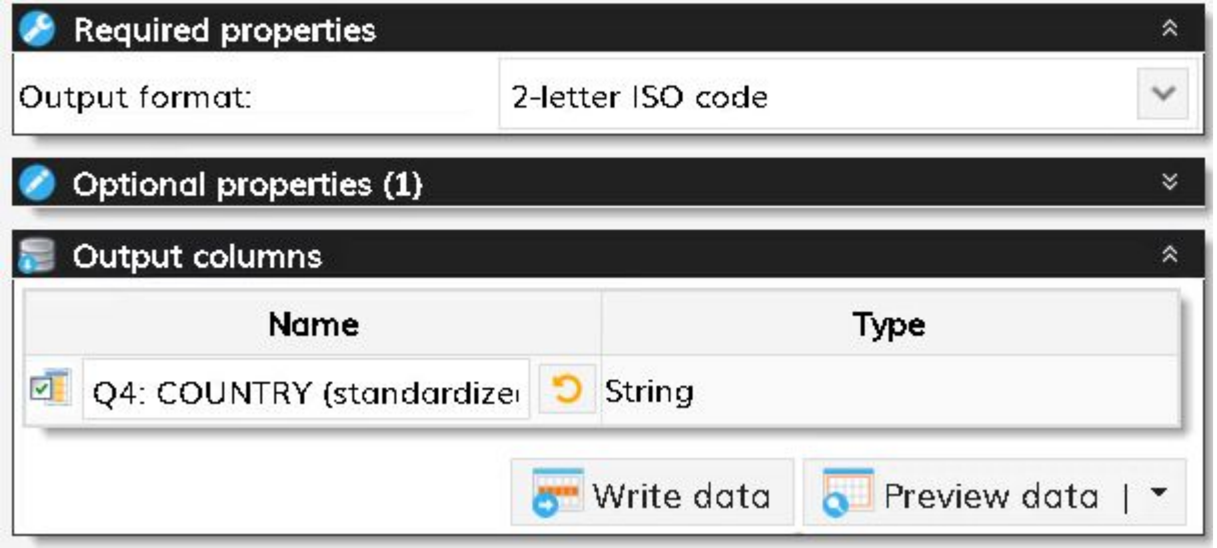




Parkgaragen.csv	ParkRide.csv
Select none	Select none
☑ XCoord	☑ XCoord
☑ YCoord	☑ YCoord
☑ OBJECTID	☑ OBJECTID
☑ NAME	☑ NAME
☑ ANSCHRIFT	☑ ANSCHRIFT
☑ ORT	☑ ORT
☑ KAT3	☑ KAT3
☑ HERKUNFT	☑ HERKUNFT
☑ PHI	☑ PHI
☑ LAMBDA	☑ LAMBDA

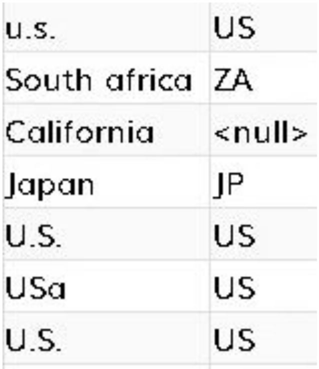
DataCleaner - Example 2: Standardization





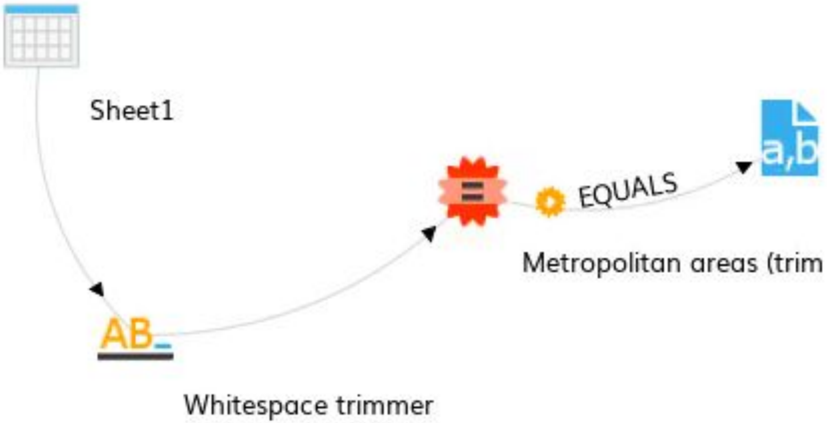
Output format: 2-letter ISO code

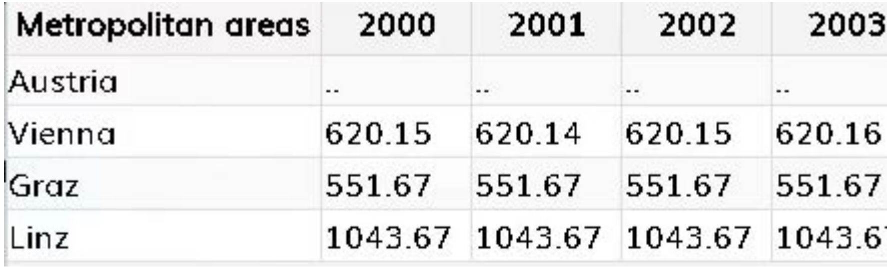
Name	Type
Q4: COUNTRY (standardize	String



u.s.	US
South africa	ZA
California	<null>
Japan	JP
U.S.	US
USa	US
U.S.	US

DataCleaner - Example 3: Filtering





Metropolitan areas	2000	2001	2002	2003
Austria	..	..	..	..
Vienna	620.15	620.14	620.15	620.16
Graz	551.67	551.67	551.67	551.67
Linz	1043.67	1043.67	1043.67	1043.6

Conclusion

Summary

Recommendation

Conclusion

- Achieve high-quality data using data cleaning tools.
- Different use cases call for different tools.
    - E.g. data analysis (DataCleaner), cleaning, transformation, (OpenRefine/Trifacta)…
- Different user requirements call for different tools.
    - E.g. data privacy (non-online tools), platform (cross-platform tools), input formats, enterprise/private use (paid vs free), …
- Some tools cater to almost all requirements. (OpenRefine)
- Others offer a subset. (Trifacta, Alteryx Designer, DataCleaner, …)
- Look at feature matrix for quick comparison according to needs.

Recommendation

Tool	Rating	Limitations
OpenRefine	+++	
Trifacta	++	Online only, paid
Alteryx Designer	+	Windows only, paid
DataCleaner	-	Breaks, unintuitive

Honorable mentions:

- Tabula (PDF data extraction) ++
- Potter's Wheel (Pioneer) -

Additional videos: Alteryx Designer, Tabula, Potter's Wheel

Thank you for your attention.