

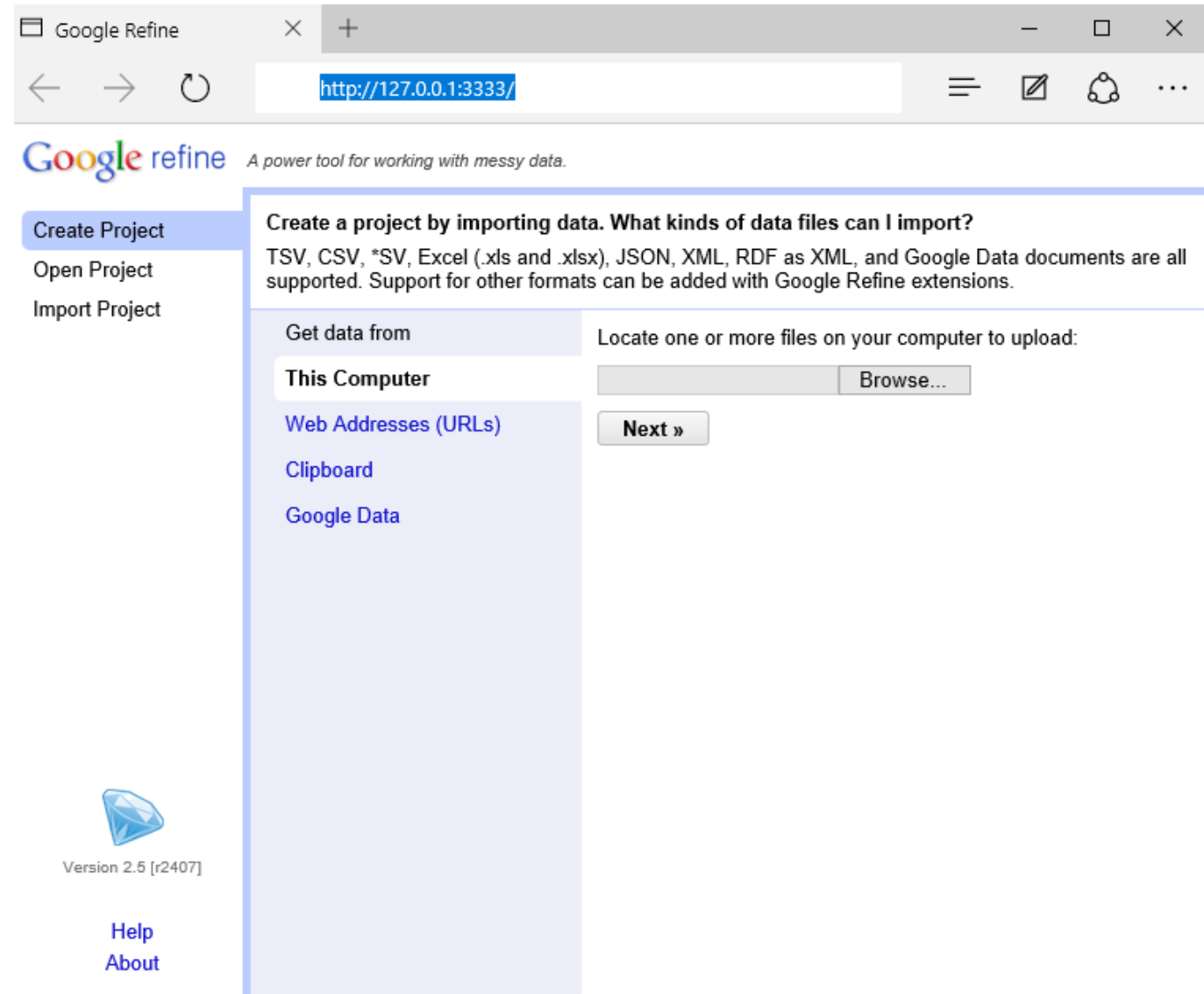


CLEANING DATA

With OpenRefine

CREATE PROJECT

- Under the “This Computer” tab, load your data from your computer into Open Refine.
- Once the data is parsed, click “Create Project” in the upper right corner to finish importing your data




The screenshot shows a web browser window titled "Google Refine" with the address bar containing "http://127.0.0.1:3333/". The page header includes the "Google refine" logo and the tagline "A power tool for working with messy data." A left sidebar contains three menu items: "Create Project" (highlighted), "Open Project", and "Import Project". The main content area is titled "Create a project by importing data. What kinds of data files can I import?" and lists supported formats: TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents. Below this, a section titled "Get data from" offers three options: "This Computer" (selected), "Web Addresses (URLs)", and "Google Data". To the right of "This Computer" is a "Browse..." button. Below the "Get data from" section is a "Next »" button. At the bottom left, there is a blue diamond icon, the text "Version 2.5 [r2407]", and links for "Help" and "About".

Facet / Filter Undo / Redo 0

4721 rows Extensions: Freebase

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

Using facets and filters



Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

All	SHIP'S NAME	AKA	SHIP'S OWNER	VESSEL TYPE	YEAR BUILT	WHERE BUILT	DATE LOST	YEAR	MNTH	DAY	LOCATION LOST
☆	1. 74 H	93015		Scow							dy Hook SE 5
☆	2. A B Crosby			Schooner	1884						ntic Highlands
☆	3. A B Sherman			Schooner	1883						e Fathoms Bank
☆	4. A B Thompson			Yacht							nt
☆	5. A C Austin			Schooner							econ
☆	6. A C Wescoat (1989)	\	A C Wescoat Co	Deck barge							dy Hook Romer
☆	7. A C Wescoat (2006)			Deck barge			7/26/2006	2006			al
☆	8. A D Scull			Schooner	1864		4/18/1880	1880	4	18	Townsend's Inlet
☆	9. A E Douglass			Schooner	1855		6/4/1862	1862	6	4	Barnegat N 8 mi
☆	10. A F Baillie			Schooner	1872		11/17/1875	1875	11	17	Carters Bar, VA

- Text facet
- Numeric facet
- Timeline facet
- Scatterplot facet
- Custom text facet...
- Custom numeric facet...
- Customized facets

- Facet
- Text filter
- Edit cells
- Edit column
- Transpose
- Sort...
- View
- Reconcile

EXPLORE YOUR DATA

Using Facets and Filters

127.0.0.1:3333/project?project=1741742654291

google refine Shipwreck Database 093015 f.xlsx Permalink

Open... Export... He

Extensions: Freebase

38 matching rows (4721 total)

Show as: rows records Show: 5 10 25 50 rows

« first < previous 1 - 10 next > last

Location Lost: change invert reset

8 choices Sort by: name count Cluster

Absecon 38 exclude

All	SHIP'S NAME	AKA	SHIP'S OWNER	VESSEL TYPE	YEAR BUILT	WHERE BUILT	DATE LOST	YEAR	MNTH	DAY	LOCATION
4.	A B Thompson			Yacht			7/18/1874	1874	7	18	Absecon
49.	Absecon			Steamship	1884	Bordentown, NJ	11/11/1885	1885	11	11	Absecon
				ooner	1863	Cleveland, OH	3/29/1890	1890	3	29	Absecon
				ooner	1881	Camden, NJ	9/6/1894	1894	9	6	Absecon
				ooner			11/14/1870	1870	11	14	Absecon
				ooner			12/24/1849	1849	12	24	Absecon
748.	Celina		Adam Hitchcock	Schooner 3 masted	1882	Bath, ME	5/12/1886	1886	5	12	Absecon
1083.	Deborah			Schooner			1/21/1855	1855	1	21	Absecon
1151.	E A Conklin		J Latham	Schooner 3 masted	1856	Port Jefferson, NY	7/28/1871	1871	7	28	Absecon
1274.	Eliza J Raynor			Schooner	1857	Cold Spring, NJ	7/1/1877	1877	7	1	Absecon

Absecon

Apply Cancel

Enter Esc

FILTER AND EDIT

Review and organize your data

Facet / Filter Undo / Redo

38 matching rows (4721 total) Extensions: Freebase

Refresh Reset All Remove All

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

LOCATION LOST change invert reset

1868 choices Sort by: name count Cluster

- Absecon 38 exclude
- Absecon bar 19
- Absecon beach 1
- Absecon Beach 41
- Absecon Beach E 12 mi 3
- Absecon Beach North Point 1
- Absecon Beach S 2 mi 1
- Absecon E 1
- Absecon E 10 mi 1
- Absecon E 35 mi 1
- Absecon E 7 mi 1
- Absecon E 9 mi 1

All	SHIP'S NAME	AKA	SHIP'S OWNER	VESSEL TYPE	YEAR BUILT	WHERE BUILT	DATE LOST	YEAR	MNTH	DAY	LOCATION LOST
☆	4.	A B Thompson		Yacht			7/18/1874	1874			econ
☆	49.	Absecon		Steamship	1884	Bordentown, NJ	11/11/1885	1885			econ
☆	149.	Alice B (1890)		Schooner	1863	Cleveland, OH					econ
☆	255.	Ann J Trainer	H W Derickson	Schooner							econ
☆	523.	Black Duck		Schooner							econ
☆	591.	Brookhaven		Schooner							econ
☆	748.	Celina	Adam Hitchcock	Schooner 3							econ
☆	1083.	Deborah		Schooner							econ
☆	1151.	E A Conklin	J Latham	Schooner 3							econ
☆	1274.	Eliza J Raynor		Schooner							econ

Facet

Text filter

Transform...

- Trim leading and trailing whitespace
- Collapse consecutive whitespace
- Unescape HTML entities
- To titlecase
- To uppercase
- To lowercase
- To number
- To date
- To text
- Blank out cells

Common transforms

- Fill down
- Blank down
- Split multi-valued cells...
- Join multi-valued cells...
- Cluster and edit...

Edit cells

Edit column

Transpose

Sort...

View

Reconcile

TRIMMING THE EXTRA

Quickly remove whitespace

FIND THE BIGGEST GROUP

Switch from “name” to “count” to find the biggest and smallest groups.

Facet / Filter

Undo / Redo 0

Refresh

Reset All

Remove All

✕ LOCATION LOST

change invert reset

1868 choices Sort by: name count

Cluster

Sandy Hook 243

Barnegat 161

Cape May 159

Long Branch, NJ 69

Squan Beach 65

Sandy Hook Romer Shoal 60

Brigantine Shoal 58

Atlantic City 50

New Jersey coast 49

Barnegat Shoal 42

Absecon Beach 41

Little Egg Harbor 41

Facet / Filter Undo / Redo

Refresh Rese

LOCATION LOST

1868 choices Sort by: name co

- Absecon 38
- Absecon bar 19
- Absecon beach 1
- Absecon Beach 41
- Absecon Beach E 12 mi 3
- Absecon Beach North Point 1
- Absecon Beach S 2 mi 1
- Absecon E 1
- Absecon E 10 mi 1
- Absecon E 35 mi 1
- Absecon E 7 mi 1
- Absecon E 9 mi 1

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method key collision Keying Function ngram-fingerprint Ngram Size 2 51 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	3	<ul style="list-style-type: none"> Cape May LSS E .25 mi (1 rows) Cape May LSS SE 2.5 mi (1 rows) Cape May LSS SSE 2.5 mi (1 rows) 	<input type="checkbox"/>	Cape May LSS E .25 mi
3	3	<ul style="list-style-type: none"> Cape May LSS SSW .75 mi (1 rows) Cape May LSS SW .75 mi (1 rows) Cape May LSS W .75 mi (1 rows) 	<input type="checkbox"/>	Cape May LSS SSW .75 mi
3	3	<ul style="list-style-type: none"> Atlantic City LSS E 1 mi (1 rows) Atlantic City LSS SE 1 mi (1 rows) Atlantic City LSS SSE 1 mi (1 rows) 	<input type="checkbox"/>	Atlantic City LSS E 1 mi
2	4	<ul style="list-style-type: none"> Avalon LSS E 1 mi (3 rows) Avalon LSS SE 1 mi (1 rows) 	<input type="checkbox"/>	Avalon LSS E 1 mi
2	2	<ul style="list-style-type: none"> Absecon Inlet E .25 mi (1 rows) Absecon Inlet E 2.5 mi (1 rows) 	<input type="checkbox"/>	Absecon Inlet E .25 mi
2	3	<ul style="list-style-type: none"> Cape May McCries Shoal (2 rows) Cape May McCries Shoal (1 rows) 	<input type="checkbox"/>	Cape May McCries Shoal

Choices in Cluster

2 — 3

Rows in Cluster

2 — 44

Average Length of Choices

11 — 28

Length Variance of Choices

0 — 1

Extensions: Freebase

previous 1 - 10 next last

DAY	LOCATION LOS
28	Sandy Hook SE 5 mi
12	Atlantic Highlands
8	Five Fathoms Bank Light
18	Absecon
	Sandy Hook Romer Shoal
11	Atlantic City Artificial Reef
25	Ocean City Artificial Reef
18	Townsend's Inlet
4	Barnegat N 8 mi
17	Carters Bar, VA

USE THE CLUSTER FUNCTION

Batch edit "like" records

<https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth>

MAKE MISTAKES!

The “Undo/Redo” Tab tracks your work so you can go back to a previous version of the data if you need to make changes or correct errors.

NOTE: You have to use Chrome to use this feature

Facet / Filter

Undo / Redo 4

Extract...

Apply...

Filter:

0. Create project

1. Mass edit 2 cells in column LOCATION
LOST

2. Mass edit 153 cells in column LOCATION
LOST

3. Mass edit 19 cells in column LOCATION
LOST

4. Mass edit 38 cells in column LOCATION
LOST

WORKING WITH NUMBERS: USE NUMERIC FACET

Google refine Shipwreck Database 093015 f.xlsx [Permalink](#)

Open... Export Help

Facet / Filter Undo / Redo

4721 rows

Extensions: [Freebase](#)

Show as: rows records Show: 5 10 25 50 rows

« first < previous 1 - 10 next > last »

Using facets and filters



Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

ION LOS	LATITUDE LOS	LONGITUDE LOS	CAUSE OF LOS	CONSTRUCTION	FLAG	LENGTH	BEAM	DRAFT	GROSS TONNA	NET TONNAGE	HOME (H
SE 5 mi			Grounded								
lands			Grounded in snowstorm		England						Aspinwall, PA
is Bank			Storm	Wood	US				612	581	Taunton, MA
			Capsized		US						Atlantic City, N
Romer			Grounded	Wood	US						
Artificial	39-15-540 N	74-14-691 W	Opened hull	Steel	US						NJ AR
Artificial	39-09-819 N	74-34-310 W	Scuttled	Steel	US						NJ AR
Inlet			Grounded	Wood	US				395	375	Great Egg Ha
8 mi			Foundered	Wood	US			9		150	Middletown, C
VA			Storm	Wood	US					284	Tuckerton, NJ

- Facet
 - Text facet
 - Numeric facet
 - Timeline facet
 - Scatterplot facet
 - Text filter
 - Edit cells
 - Edit column
 - Transpose
 - Sort..
 - View
 - Reconcile
- Custom text facet..
 - Custom numeric facet..
 - Customized facets


Facet / Filter Undo / Redo 0

Refresh Reset All Remove All

4721 rows

LENGTH change reset

grel:value



10.00 — 810.00

Numeric 1757 Non-numeric 431 Blank 2533 Error 0

Edit Facet's Expression based on Column LENGTH

Expression Language Google Refine Expression Language (GREL) ▾

value.log() No syntax error.

Preview History Starred Help

row	value	value.log()
1.	null	Error: log expects a number
2.	null	Error: log expects a number
3.	154.5	2.1889284837608534
4.	null	Error: log expects a number
5.	null	Error: log expects a number
6.	60	Error: log expects a number
7.	60	Error: log expects a number


OK Cancel

Facet / Filter Undo / Redo 0

Refresh Reset All Remove All

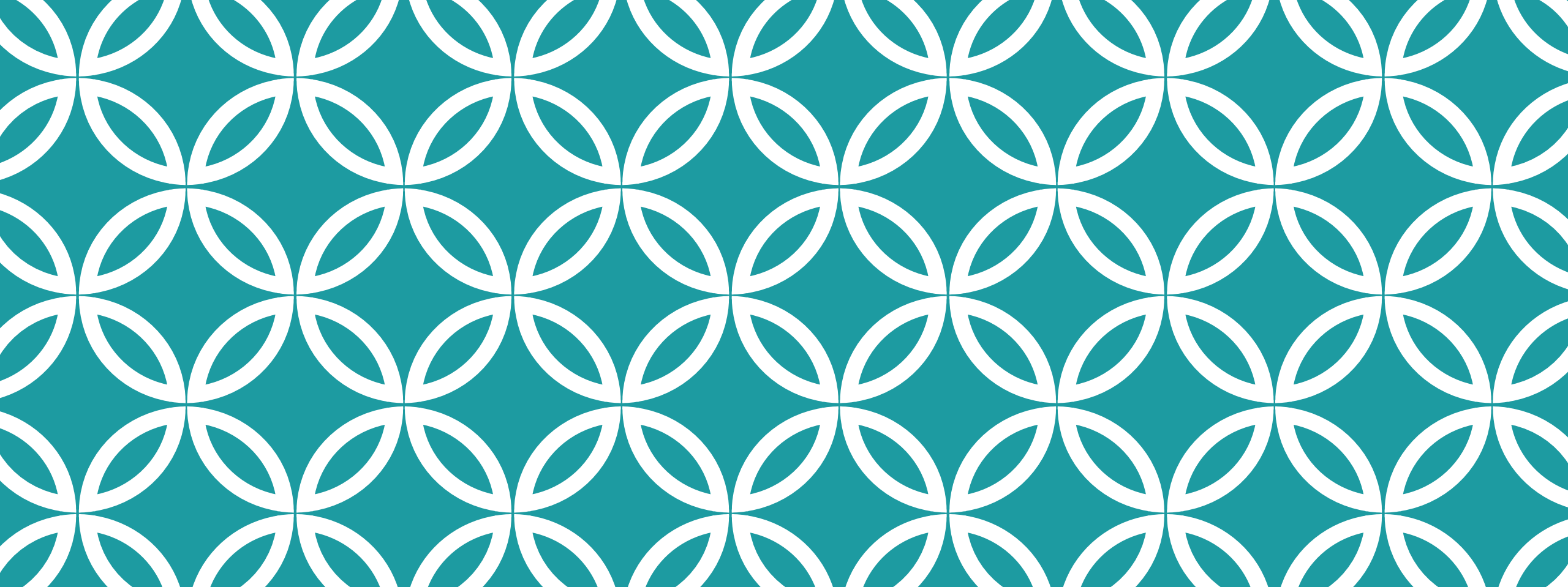
LENGTH change reset

grel:value.log()



1.29 — 2.91

Numeric 1757 Non-numeric 0 Blank 0 Error 2964



DIG IN

Dive into your data!