

AmPLY  
Innovations  
Private Limited

# Open Refine Task

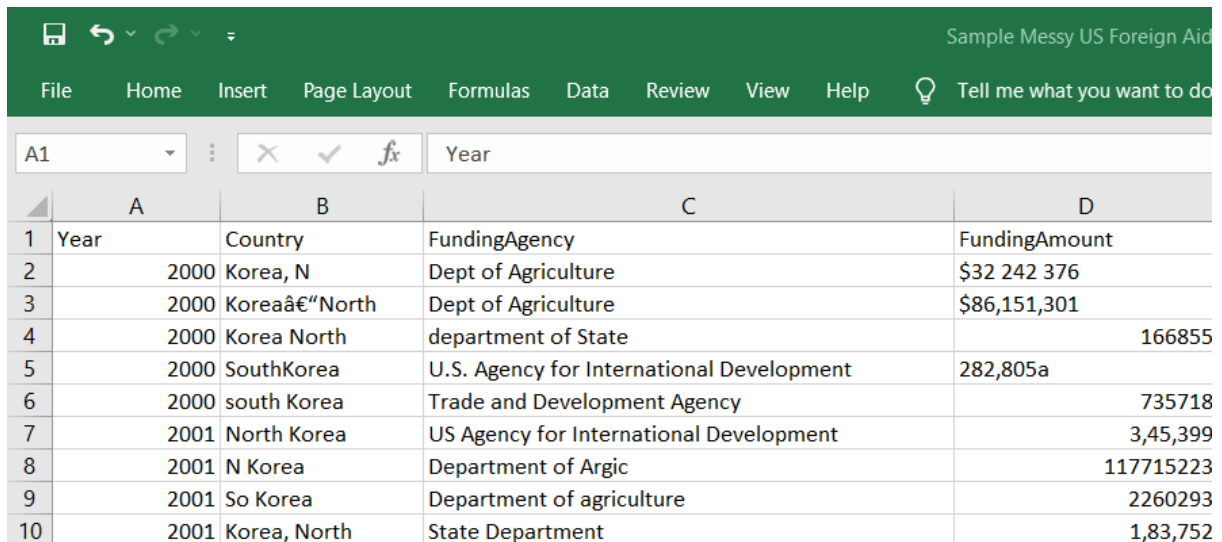
Submitted by:  
Aneesa Begum J

---

# Open Refine Project

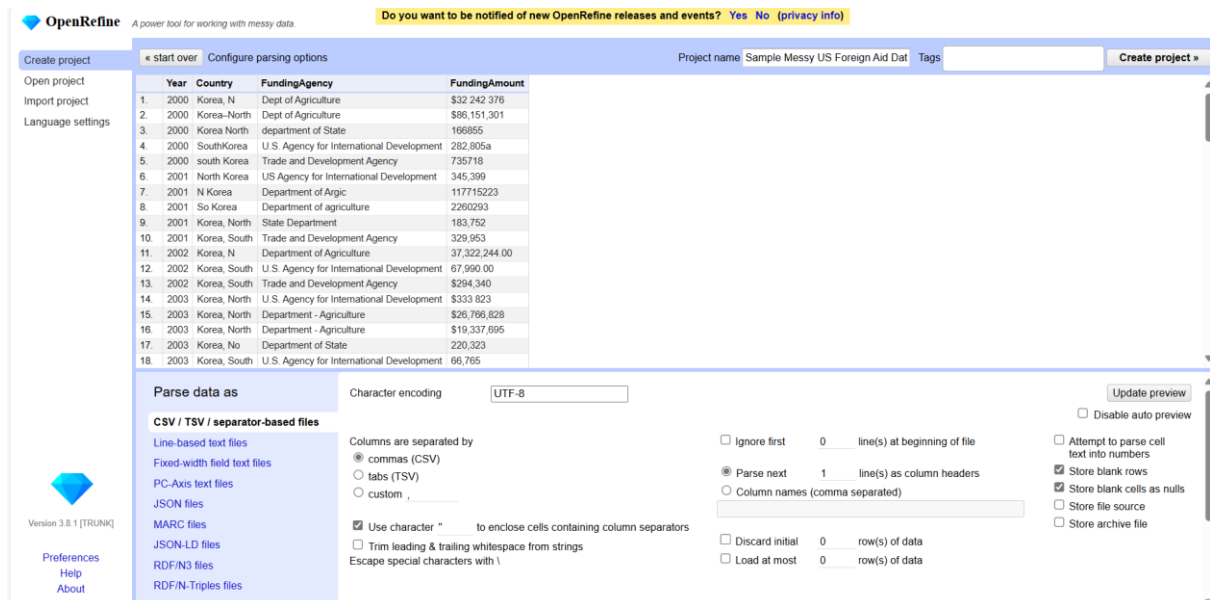
**Task :** Data cleaning and pre-processing

**Dataset used:** Sample Messy US Foreign AID Dataset



	A	B	C	D
1	Year	Country	FundingAgency	FundingAmount
2	2000	Korea, N	Dept of Agriculture	\$32 242 376
3	2000	Koreaâ€œNorth	Dept of Agriculture	\$86,151,301
4	2000	Korea North	department of State	166855
5	2000	SouthKorea	U.S. Agency for International Development	282,805a
6	2000	south Korea	Trade and Development Agency	735718
7	2001	North Korea	US Agency for International Development	3,45,399
8	2001	N Korea	Department of Argic	117715223
9	2001	So Korea	Department of agriculture	2260293
10	2001	Korea, North	State Department	1,83,752

**Step 1:** Load the Dataset in Open Refine



OpenRefine A power tool for working with messy data. Do you want to be notified of new OpenRefine releases and events? Yes No (privacy info)

Create project Open project Import project Language settings

Project name Sample Messy US Foreign Aid Dat Tags Create project »

Parse data as

Character encoding UTF-8

Update preview

Disable auto preview

Columns are separated by

Ignore first 0 line(s) at beginning of file

Parse next 1 line(s) as column headers

Column names (comma separated)

Discard initial 0 row(s) of data

Load at most 0 row(s) of data

Attempt to parse cell text into numbers

Store blank rows

Store blank cells as nulls

Store file source

Store archive file

Version 3.8.1 [TRUNK]

Preferences Help About

**Step 2:** Create a project



Project name Sample Messy US Foreign Aid Dat Tags Create project »

### Step 3: Standardize the Data

- Convert Dollar Amounts from text to numbers in Funding Amount Column

127 rows						
Show as: <b>rows</b> records			Show: 5 10 25 50 100 500 1000 rows			
All	Year	Country	FundingAgency	FundingAmount		
★	1.	2000	Korea, N	Dept of Agriculture	\$32 242 376	
★	2.	2000	Korea–North	Dept of Agriculture	\$86,151,301	
★	3.	2000	Korea North	department of State	166855	
★	4.	2000	SouthKorea	U.S. Agency for International Development	282,805a	
★	5.	2000	south Korea	Trade and Development Agency	735718	
★	6.	2001	North Korea	US Agency for International Development	345,399	
★	7.	2001	N Korea	Department of Argic	117715223	
★	8.	2001	So Korea	Department of agriculture	2260293	
★	9.	2001	Korea, North	State Department	183,752	
★	10.	2001	Korea, South	Trade and Development Agency	329,953	

Some of the rows that are colored green are transformed into numbers. But, the 1<sup>st</sup> two rows and most of the other rows are not transformed.

- Removing \$ and , from the Funding Amount column.

### Custom text transform on column FundingAmount

Expression Language General Refine Expression Language (GREL) ▾

`value.replace(',', '')` No syntax error.

Preview			History	Starred	Help
row	value	value.replace(',', '')			
1.	\$32 242 376	\$32 242 376			
2.	\$86,151,301	\$86151301			
3.	166855	166855.0			
4.	282,805a	282805a			
5.	735718	735718.0			
6.	345,399	345399			

## Custom text transform on column FundingAmount

Expression

Language

General Refine Expression Language (GREL) ▼

value.replace('\$', '')

No syntax error.

Preview

History

Starred

Help

row	value	value.replace('\$', '')
1.	\$32 242 376	32 242 376
2.	\$86151301	86151301
3.	166855.0	166855.0
4.	282805a	282805a
5.	735718.0	735718.0
6.	345399	345399

127 rows

Show as: **rows** records Show: 5 10 25 50 100 500 1000 rows

▼ All			▼ Year	▼ Country	▼ FundingAgency	▼ FundingAmount
☆	🚩	1.	2000	Korea, N	Dept of Agriculture	32242376
☆	🚩	2.	2000	Korea—North	Dept of Agriculture	86151301
☆	🚩	3.	2000	Korea North	department of State	166855
☆	🚩	4.	2000	SouthKorea	U.S. Agency for International Development	282805a
☆	🚩	5.	2000	south Korea	Trade and Development Agency	735718
☆	🚩	6.	2001	North Korea	US Agency for International Development	345399
☆	🚩	7.	2001	N Korea	Department of Argic	117715223
☆	🚩	8.	2001	So Korea	Department of agriculture	2260293
☆	🚩	9.	2001	Korea, North	State Department	183752
☆	🚩	10.	2001	Korea, South	Trade and Development Agency	329953
☆	🚩	11.	2002	Korea, N	Department of Agriculture	37322244
☆	🚩	12.	2002	Korea, South	U.S. Agency for International Development	67990
☆	🚩	13.	2002	Korea, South	Trade and Development Agency	294340
☆	🚩	14.	2003	Korea, North	U.S. Agency for International Development	333 823
☆	🚩	15.	2003	Korea, North	Department - Agriculture	26766828

Still few rows have non-numeric rows with the space and ending letter 'a'.

- Fixing it manually by clicking 'edit' on each non-numeric row

FundingAmount	
32242376	
86151301	
166855	
282805a	edit
735718	
345399	

FundingAmount	Data type: <input type="text" value="number"/>
32242376	<input type="text" value="32242376"/>
<input type="button" value="Apply"/> <input type="button" value="Apply to all identical cells"/> <input type="button" value="Cancel"/>	Enter      Ctrl-Enter      Esc

FundingAmount	Data type: <input type="text" value="number"/>
282805a	<input type="text" value="282805"/>
<input type="button" value="Apply"/> <input type="button" value="Apply to all identical cells"/> <input type="button" value="Cancel"/>	Enter      Ctrl-Enter      Esc

#### Step 4: Checking for null values

127 rows					
Show as: <b>rows</b> records		Show: 5 10 25 50 100 500 1000 rows			
« first		< previous		1	
All	Year	Country	FundingAgency	FundingAmount	
1.	2000	Korea, N	Dept of Agriculture	Facet	Text facet
2.	2000	Korea--North	Dept of Agriculture	Text filter	Numeric facet
3.	2000	Korea North	department of State	Edit cells	Timeline facet
4.	2000	SouthKorea	U.S. Agency for International Development	Edit column	Scatterplot facet...
5.	2000	south Korea	Trade and Development Agency	Transpose	Custom text facet...
6.	2001	North Korea	US Agency for International Development	Sort...	Custom numeric facet...
7.	2001	N Korea	Department of Argic	View	Customized facets
8.	2001	So Korea	Department of agriculture	Reconcile	Word facet
9.	2001	Korea, North	State Department		Duplicates facet
10.	2001	Korea, South	Trade and Development Agency		Numeric log facet
11.	2002	Korea, N	Department of Agriculture	37322244	1-bounded numeric log facet
12.	2002	Korea, South	U.S. Agency for International Development	67990	Text length facet
13.	2002	Korea, South	Trade and Development Agency	294340	Log of text length facet
14.	2003	Korea, North	U.S. Agency for International Development	333823	Unicode char-code facet
15.	2003	Korea, North	Department - Agriculture	26766828	Facet by error
16.	2003	Korea, North	Department - Agriculture	19337695	Facet by null
17.	2003	Korea, No	Department of State	220323	Facet by empty string
18.	2003	Korea, South	U.S. Agency for International Development	66765	Facet by blank (null or empty string)
19.	2003	Korea, South	Trade and Development Agency	19899	
20.	2004	Korea, North	U.S. Agency f/ International Development	782473	
21.	2004	Korea, North	U.S. Agency: International Development	311432	
22.	2004	Korea, North	U.S. Agency for International Development	460355	

Facet / Filter
Undo / Redo 7 / 7

Refresh
Reset all
Remove all

x
FundingAmount
change

1 choice Sort by: name count

false 127

Facet by choice counts

There are no null values found.

## Step 5: Checking for Duplicates

127 rows

Show as: rows records Show: 5 10 25 50 100 500 1000 rows « first < previous

All	Year	Country	FundingAgency	FundingAmount	
1.	2000	Korea, N	Dept of Agriculture		Facet
2.	2000	Korea-North	Dept of Agriculture		Text filter
3.	2000	Korea North	department of State		Edit cells
4.	2000	SouthKorea	U.S. Agency for International Development		Edit column
5.	2000	south Korea	Trade and Development Agency		Transpose
6.	2001	North Korea	US Agency for International Development		Sort...
7.	2001	N Korea	Department of Argic		View
8.	2001	So Korea	Department of agriculture		Reconcile
9.	2001	Korea, North	State Department		
10.	2001	Korea, South	Trade and Development Agency		
11.	2002	Korea, N	Department of Agriculture	37322244	
12.	2002	Korea, South	U.S. Agency for International Development	67990	
13.	2002	Korea, South	Trade and Development Agency	294340	
14.	2003	Korea, North	U.S. Agency for International Development	333823	
15.	2003	Korea, North	Department - Agriculture	26766828	
16.	2003	Korea, North	Department - Agriculture	19337695	
17.	2003	Korea, No	Department of State	220323	
18.	2003	Korea, South	U.S. Agency for International Development	66765	
19.	2003	Korea, South	Trade and Development Agency	19899	
20.	2004	Korea, North	U.S. Agency f/ International Development	782473	
21.	2004	Korea, North	U.S. Agency: International Development	311432	

Text facet  
Numeric facet  
Timeline facet  
Scatterplot facet...  
Custom text facet...  
Custom numeric facet...  
Customized facets  
Word facet  
Duplicates facet  
Numeric log facet  
1-bounded numeric log facet  
Text length facet  
Log of text length facet  
Unicode char-code facet  
Facet by error  
Facet by null  
Facet by empty string  
Facet by blank (null or empty string)

x
FundingAmount
change

1 choice Sort by: name count

false 127

Facet by choice counts

There are no duplicates found on this dataset.

127 rows

Show as: **rows** records Show: 5 10 25 50 100 500 1000 rows

▼ All			▼ Year	▼ Country	▼ FundingAgency	▼ FundingAmount
☆	🔊	1.	2000	Korea, N	Dept of Agriculture	32242376
☆	🔊	2.	2000	Korea–North	Dept of Agriculture	86151301
☆	🔊	3.	2000	Korea North	department of State	166855
☆	🔊	4.	2000	SouthKorea	U.S. Agency for International Development	282805
☆	🔊	5.	2000	south K <a href="#">edit</a>	Trade and Development Agency	735718
☆	🔊	6.	2001	North Korea	US Agency for International Development	345399
☆	🔊	7.	2001	N Korea	Department of Argic	117715223
☆	🔊	8.	2001	So Korea	Department of agriculture	2260293
☆	🔊	9.	2001	Korea, North	State Department	183752
☆	🔊	10.	2001	Korea, South	Trade and Development Agency	329953
☆	🔊	11.	2002	Korea, N	Department of Agriculture	37322244
☆	🔊	12.	2002	Korea, South	U.S. Agency for International Development	67990
☆	🔊	13.	2002	Korea, South	Trade and Development Agency	294340
☆	🔊	14.	2003	Korea, North	U.S. Agency for International Development	333823
☆	🔊	15.	2003	Korea, North	Department - Agriculture	26766828
☆	🔊	16.	2003	Korea, North	Department - Agriculture	19337695
☆	🔊	17.	2003	Korea, No	Department of State	220323
☆	🔊	18.	2003	Korea, South	U.S. Agency for International Development	66765
☆	🔊	19.	2003	Korea, South	Trade and Development Agency	19899
☆	🔊	20.	2004	Korea, North	U.S. Agency f/ International Development	782473
☆	🔊	21.	2004	Korea, North	U.S. Agency: International Development	311432
☆	🔊	22.	2004	Korea, North	U.S. Agency for International Development	460355
☆	🔊	23.	2004	Korea, North	Department of Agriculture	11503280
☆	🔊	24.	2004	Korea, N	Department of Agriculture	54680490
☆	🔊	25.	2004	Korea, North	Department of State	1108637

Therefore, all the rows in the Funding Amount column is cleaned and are converted into the numeric values.

## Step 6: Cluster Similar Spelling Vaues

**Cluster and edit column "Country"**

Find groups of different cell values that might be other representations of the same thing. For example, "New York" and "new york" likely refer to the same concept and just differ by capitalization, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method: Key collision Keying function: Fingerprint ☒ Auto-update 7 clusters found

Cluster size	Row count	Values in cluster	Merge?	New cell value
7	57	<ul style="list-style-type: none"><li>Korea, South (49 rows)</li><li>South Korea (3 rows)</li><li>Korea South</li><li>Korea `South</li><li>Korea: South</li><li>south Korea</li><li>south korea</li></ul>	<input checked="" type="checkbox"/>	Korea, South
5	51	<ul style="list-style-type: none"><li>Korea, North (47 rows)</li><li>Korea ;North</li><li>Korea North</li><li>Korea north</li><li>North Korea</li></ul>	<input checked="" type="checkbox"/>	Korea, North
3	3	<ul style="list-style-type: none"><li>Korea-North</li><li>Korea:North</li><li>Korea-North</li></ul>	<input checked="" type="checkbox"/>	Korea-North
2	5	<ul style="list-style-type: none"><li>Korea, N (4 rows)</li><li>N Korea</li></ul>	<input checked="" type="checkbox"/>	Korea, N
2	3	<ul style="list-style-type: none"><li>Korea, No (2 rows)</li><li>Korea. No</li></ul>	<input checked="" type="checkbox"/>	Korea, No

# Choices in cluster: 2 — 7

# Rows in cluster: 2 — 57

Average length of choices: 7.5 — 11.5

Length variance of choices: 0.47100000000000003 — 1

Select all Deselect all Export clusters Merge selected & re-cluster Merge selected & Close Close

Therefore, the Sample Messy US Foreign AID Dataset is clean now.

## Step 7: Checking the numeric facet of the column "Year"

All	Year	Country	FundingAgency	FundingAmount
1.	Facet		Text facet	32242376
2.	Text filter		Numeric facet	86151301
3.	Edit cells		Timeline facet	166855
4.	Edit column		Scatterplot facet...	282805
5.	Transpose		Custom text facet...	735718
6.	Sort...		Custom numeric facet...	345399
7.	View		Customized facets	117715223
8.	Reconcile		State Department	2260293
9.			Trade and Development Agency	183752
10.				329953

